STATISTICAL METHODS FOR STUDENTS IN EDUCATION

HOLZINGER

To the Krengers

From

Hattie Francis Smith

used in my course -

KARL JOHN HOLZINGER 1954

To Keep LMR

Professor Karl John Holzinger unexpectedly died in his home on January 15. Professor Holzinger, who was sixty-one years of age, had been connected with the Department of Education for thirty-two years. With the exception of time taken out to study with Karl Pearson and Charles Spearman at the University of London in 1922 and 1925, for service in the Psychological Corps of the United States Army in World War I, and for short teaching assignments in the University of California (Berkeley) and in other institutions, Professor Holzinger spent his entire professional life on the campus of the University of Chicago.

From the time he received the Ph.D. summa cum laude from the University in 1922 until his death, Mr. Holzinger was a national and an international figure in the field of statistics. Many of his former students hold key positions in Education and in government.

Professor Holzinger was a part president of the Psychometric Society and past vice-president of the American Statistical Association. He was author of many books and monographs. Best known among them are Statistical Methods; Influence of Environment on Intelligence, Achievement, and Conduct (with F. N. Freeman); Factor Analysis: A Synthesis (with H. Harman); Twins (with Freeman and Newman); A Study in Factor Analysis: The Reliability of Bi-factors and their Relation to other Measures (with Frances Swineford); and The Spearman-Holzinger Preliminary Reports (1-9) on Factor Analysis. Professor Holzinger was joint editor of the Journal of Educational Psychology from 1949 to the time of his death.

STATISTICAL METHODS FOR STUDENTS IN EDUCATION

BY

KARL J. HOLZINGER

ASSOCIATE PROFESSOR OF EDUCATION, THE UNIVERSITY OF CHICAGO

GINN AND COMPANY

BOSTON · NEW YORK · CHICAGO · LONDON

ATLANTA · DALLAS · COLUMBUS · SAN FRANCISCO

COPYRIGHT, 1928, BY KARL J. HOLZINGER
ALL RIGHTS RESERVED

PRINTED IN THE UNITED STATES OF AMERICA

628.1

The Athenæum Press
GINN AND COMPANY • PROPRIETORS • BOSTON • U.S.A.

PREFACE

This book is intended for students in education, who usually have had little training in mathematics. For those who have had considerable mathematics the theory of statistics is comparatively easy, but for students without such training the more advanced statistical methods offer many difficulties. The present volume supplements the mathematical preparation of the student by including sections on such topics as graphing, logarithms, and elementary theory of probability. The proofs of difficult theorems have been omitted throughout and demonstrations have been included only when experience has shown that they come within the grasp of the ordinary student and assist in a clear understanding of the method involved.

Although no attempt has been made to include all statistical methods now used in the field of education, the present text treats a somewhat larger number than will be found in most elementary books. The chief additions to the usual topics are the percentile method, application of the normal curve in correlating qualitative series, partial and multiple correlation, and elementary theory of curve fitting. The important subject of index numbers has been omitted entirely because a satisfactory treatment is beyond the scope of this book. The increasing need for index numbers in the field of school costs will probably lead to a separate volume on these methods.

In order to insure a clear understanding of the statistical arithmetic involved in the various methods presented, complete model problems have been worked out in the text. The experience of the writer has been that the ordinary student has considerable difficulty in formulating his plans for calculation and is greatly assisted by detailed arithmetical schemes

for computation, particularly in the early part of the course. A considerable number of exercises with answers have been added at the end of each chapter to clarify the methods discussed and to afford the student sufficient arithmetical practice to enable him to become accurate in his work. The amount of such practice needed varies greatly with students, and enough exercises are included to meet the needs of those requiring most drill.

The material in this volume will be found sufficient for an ordinary course of six months, but it may be condensed for a shorter course by the omission of certain topics and chapters. For an introductory course in a normal school or college, Chapters I to IX with selected topics from Chapters XII, XIII, and XIV are suggested. In case a second course is offered, the last seven or eight chapters with supplementary reading and term papers will usually be ample.

The writer is greatly indebted to Professor Karl Pearson, Dr. Leonard P. Ayres, and Professor Harold Rugg for ideas acquired while he was under their instruction. Valuable advice and suggestions have also been contributed by Dr. Egon Pearson, Professor C. H. Judd, Professor F. N. Freeman, Professor E. R. Breslich, Dr. Douglas Scates, and Dr. Ralph Hogan, all of whom read the manuscript while in preparation. Additional thanks are due to Mr. Lumir Brazda for preparation of the diagrams and to Mrs. Bryan Mitchell for assistance in checking the proof.

KARL J. HOLZINGER

CONTENTS

CHAPTER PAGE

I. INTRODUCTION 1

1. The Need for Statistical Method in Dealing with Educational Problems 1
2. Some General Requirements for Success in the Use of Statistical Method 3
3. General Statistical Procedure in Dealing with a Problem . 5

II. COLLECTION AND CLASSIFICATION OF DATA 9

1. Primary and Secondary Data 9
2. Some Examples of Secondary Source Material 10
3. Units of Collection 11
4. Types of Series 12
5. Methods of Collecting Data 14
6. Sampling 16
7. Arrangement of the Original Data 19
8. The Simple Frequency Distribution 22
9. The Classifier 25
10. Cumulative Frequency Distributions 28

III. TABALAR AND GRAPHICAL PRESENTATION OF DATA . . 31

1. Purpose of Tables and Diagrams 31
2. The Construction of Tables for Presentation 32
3. Column and Bar Diagrams 36
4. Coördinates 40
5. Functional Relationships 42
6. The Straight Line 43
7. Non-Linear Relationships 44

IV. LOGARITHMS 47

1. Introductory 47
2. Arithmetical and Geometrical Progressions 47
3. The Invention of Logarithms 49
4. Laws of Exponents 51
5. Laws of Logarithms 52
6. The Briggs System of Logarithms 53
7. Interpolation 56
8. Some Additional Problems 59

CHAPTER PAGE

V. ERRORS IN CALCULATION AND MEASUREMENT 65
1. Accuracy in Statistical Method 65
2. Absolute and Relative Errors 65
3. Biased and Unbiased Errors 66
4. Significant Figures 68
5. Arithmetical Computation with Rounded Numbers . . . 69
6. Logarithmic Computation with Rounded Numbers . . . 71
7. Errors in Educational Measurement 74

VI. AVERAGES 78
1. Introductory 78
2. Calculation of the Mean 79
3. Properties of the Mean 83
4. Calculation of the Median 85
5. Properties of the Median 88
6. The Crude Mode 90
7. The Geometric Mean and Geometrical Series 91
8. The Harmonic Mean 95

VII. MEASURES OF DISPERSION 101
1. Introductory 101
2. Mean Deviation 102
3. The Standard Deviation 108
4. The Quartile Deviation 110
5. Comparison of Measures of Dispersion 113
6. The Coefficient of Variation 116
7. Comparable Measurements 118
8. The Measurement of Skewness 122

VIII. THE PERCENTILE METHOD 127
1. Introductory 127
2. Percentiles 127
3. Percentile Curves 131
4. Use of Percentile Curves 134
5. Percentile Ranks 136

IX. LINEAR CORRELATION WITH QUANTITATIVE SERIES . 141
1. The Meaning of Correlation 141
2. The Product-Moment Correlation Coefficient 143
3. The Computation of the Correlation Coefficient with Ungrouped Items 146

CHAPTER PAGE

4. The Computation of the Correlation Coefficient for a Frequency Table 149
5. Lines of Regression 154
6. The Interpretation of the Correlation Coefficient 163
7. Some Uses of Correlation in Evaluating Test Material . . 167
8. The Effect of Selection upon Correlation 172
9. The Effect of Range of Talent upon Correlation 172

X. NON-LINEAR CORRELATION 177

1. The Correlation Ratio 177
2. Modified Formulas for the Correlation Ratios 179
3. A Combination Form for the Correlation Coefficient and Ratios 181
4. Tests for Linearity 183
5. A Method of Eliminating the Effect of a Variable upon the Association between Two Others 184

XI. THE BINOMIAL DISTRIBUTION 190

1. Introductory 190
2. Permutations and Combinations 191
3. Elementary Probability 192
4. The Binomial Theorem and the Point Binomial 195
5. The Mean of the Point Binomial and its Standard Deviation 197
6. Experimental Verification of the Binomial Law 199
7. The Binomial Applied to Statistical Data 201

XII. THE NORMAL PROBABILITY CURVE 204

1. Introductory 204
2. The Equation of the Normal Probability Curve 207
3. The Area, Ordinates, and Deviates of the Normal Curve 209
4. Comparison of the Point Binomial and the Normal Curve 212
5. Fitting a Normal Curve to a Frequency Distribution of Data 214
6. Some Properties of the Normal Curve 217
7. Representing Data on a Normal Scale 221
8. The Scaling of Test Questions 224

XIII. SAMPLING AND RESPONSE ERRORS 231

1. Introductory 231
2. Sampling Error in the Mean 232
3. The Probable Error of the Difference between Two Means 235
4. The Probable Errors of Certain Constants for a Normal Distribution 237

In the selection and organization of test material and the standardization and preparation in final form, elaborate technique is often required. Modern developments in test construction have led to the use of more and more refined methods, so that the test-maker of today needs to be a thorough student of statistics.

In the application of standardized tests to such problems as pupil classification, vocational guidance, diagnosis of special abilities, and evaluation of methods of instruction, a sound knowledge of statistical method is imperative, because all such studies involve the collection of appropriate data, summarization of the results, and correct inferences from the statistical findings.

The quantitative trend in school investigation has given rise to a tremendous bulk of literature. There are now hundreds of volumes on school surveys filled with tables and diagrams; there are books, monographs, theses, and reports likewise replete with statistics; there are scores of government, state, and institutional pamphlets; there are hundreds of standardized tests; and there is an ever-increasing amount of periodical literature reporting the findings of quantitative studies.

It is evident that if the school administrators and teachers for whom a large part of this great body of literature was written are to understand and apply it, they must have considerable familiarity with statistical method. It is impossible to keep up with the most recent developments in school research without some knowledge of the methods upon which such investigations are based.

Professional schools and departments in universities devoted to the training of teachers and administrators are meeting the demand by courses in experimental and statistical method. The purpose of such courses, in general, is to give the student sufficient information for intelligent reading of the present quantitative literature, and to furnish him with the technique necessary for carrying on his own investigations. This twofold aim has been kept in mind in preparing the present text.

2. Some General Requirements for Success in the Use of Statistical Method

In conducting a statistical study the investigator, survey expert, or classroom teacher should have in mind some definite problem or purpose, no matter how limited in scope. The mere gathering of masses of data or the haphazard calculation and plotting of diagrams are of little value unless they can be brought to bear upon a problem. While desirable lines of investigation are often discovered after the data have been collected and tabulated in a tentative way, it is much safer to decide upon the problem first and then proceed to collect the data necessary for its solution. The selection of a problem which is worth while, and which is sufficiently limited so that controls may be made and all necessary details carried out thoroughly and completely, is perhaps the most difficult part of the whole statistical procedure. It requires wide knowledge of the general field in which the problem lies, and a certain *constructive imagination* in foreseeing the various difficulties which are likely to arise.

Another requisite for a good statistical investigation is adequate data. No matter how excellent the problem or the plan of procedure, if the data employed are scanty the results will be of little value. Statistical method usually involves some generalization based upon summaries of the data. If the data are small in number, therefore, the conclusions drawn will not be reliable. This may be illustrated by some unpublished experiments in maze-learning based upon about twenty-five cases. Out of eight similar studies five showed a superiority for one method of learning, while the other three showed a difference in favor of another method. In all the experiments the number of cases was so small that none of the differences obtained proved to be significant, but could be readily accounted for by mere chance fluctuations in the samples of data chosen. While there is no fixed number of cases necessary for making

a statistical study, a desirable minimum for experimental work is about fifty, provided they are well chosen.

Data adequate as to number are not alone sufficient to insure satisfactory material. The facts gathered must be reliable and pertinent to the problem in hand. Questionnaire returns often fail in this respect because the intelligent replies of a number of persons to whom the blanks are sent are offset by careless or random answers on the part of others. Increasing the bulk of such data is not likely to increase its reliability, but the selecting of even a smaller number of persons who could be depended upon to give careful replies would yield better results. Thus if one wished to discover the most important aims in the teaching of high-school English, returns from a small well-selected group of experienced teachers would be preferable to those from a much larger group taken at random.

It frequently happens that the worker loses sight of the fact that his data are inadequate as to *quantity* and *quality* and applies elaborate statistical methods with the expectation that the final results will be of value. Such procedure, if followed intentionally, has been rightly described as "hiding behind a statistical smoke-screen," and is nothing less than a scientific crime. The limitations of the data employed should always be frankly recognized and the conclusions of the study made with them in mind. No amount of subsequent juggling by complicated formulas can give good results when they are based upon originally faulty data.

The successful statistician must have the capacity for careful, painstaking, and scientifically honest work. It is so easy to gather a few figures and tabulate them in such a way as to show a desired result or "prove" a certain theory that the temptations on the path of scientific rectitude are great. The untrained reader is often so bewildered by tables and diagrams that he is incapable of verifying the method or the inferences in a statistical article and either accepts the conclusions on the reputation of the writer or perhaps concludes that "anything can be proved

by statistics." Educational science would be greatly improved by the production of a smaller number of studies based upon better data and a more cautious use of statistical method.

A final requisite for the successful use of statistics is training in methodology. The investigator needs to become familiar with the various technical methods and processes of calculation. He needs much training in the application of these methods to data and problems in the particular field in which he expects to work. He also needs some knowledge of the difficult field of statistical inference. It is this general pedagogical requirement which the textbook and course in statistics are expected to fulfill. Such a course of study should familiarize the student with methods appropriate to educational problems, insure skill in statistical arithmetic, and provide opportunity for working out a worthwhile problem under careful guidance.

3. General Statistical Procedure in Dealing with a Problem

While there is no set order in which the steps in a statistical study must be carried out, experience has shown that a systematic procedure like the following is logical and economical of time and labor. Most of these steps will be discussed and fully illustrated in subsequent chapters.

(1) **Planning of the study.** When the student has some problem selected, his first concern will be with a rough plan for the whole study. It may not be possible to define the problem very specifically until the data have been gathered and examined, but the more definitely the limits of the inquiry can be set in advance the easier will be the subsequent steps. The usual mistake is to select a problem much too broad and too difficult for any one individual or even a small group of workers to undertake effectively. The availability, sources, accuracy, and methods of gathering data should all be considered in the preliminary plan.

(2) **Collection of the data.** With the problem defined and a general plan made, the next step is to collect the necessary data. This is accomplished by the use of questionnaires, by personal tabulation from data already available in records, or by the application of standardized tests, rating schemes, and other such measuring devices (Chapter II).

(3) **Preliminary analysis of the data.** If a questionnaire has been used in collecting the material, it is usually necessary to examine the returns very carefully before making tabulations. Incompleteness, inaccuracy, and ambiguity in the answers given should all be considered before the data are used. Similar analysis is often necessary with the results of standardized scales; unusual test conditions and errors in giving and in scoring the tests need to be checked up before tabulation is begun.

A preliminary analysis of the material will also be desirable in many cases to determine whether or not the data are adequate for the problem in hand. It may be that question-blank returns from a certain source are too scanty or fail to appear in such a form as to meet the requirements of the problem. In such cases a revised blank and more data will be required (Chapter II).

(4) **Tabulation for primary records.** After the preliminary analysis the data should be tabulated in such a way as to form both a permanent and a convenient working record. The permanent record may be kept in a bound volume with a page to each case or in the form of a master sheet with the names and records in parallel columns. The working record is usually in the form of small cards. One of these is made out for each case and the data entered in compact form so that the cards may be readily sorted and the resulting distributions easily checked (section 7, Chapter II).

(5) **Classification of the material.** Distributions, tables, and serial arrangements may next be made from the primary record.

These furnish the basis for calculations and graphical representations of the material.

(6) **Analysis of the classified data and planning of the calculations.** The particular statistical calculations to be employed are often not apparent until the data have been arranged in systematic form. The choice and right use of the proper analytical methods are extremely important, and at this point sound statistical judgment is required.

After the required calculations have been decided upon, they should be planned throughout before computation is begun. This is particularly advisable with data involving correlations (Chapters IX and X), the tables for which may be checked against one another and also used to furnish other statistical quantities such as the averages and measures of variability (Chapters VI and VII).

(7) **Calculation of the statistical constants.** The computations required may be made with the assistance of calculating tables and machines. It is desirable to have complete checks on the arithmetical accuracy of the work. Some of these are afforded by formulas, but the best check is to have two persons perform the calculations independently (Chapter V).

(8) **Interpretation of results.** The study has now reached the point where a careful scrutiny of results is required. These need to be interpreted in terms of the problem in hand. If the investigator is fortunate, the results may come out in such a way that the conclusions to be drawn are clear-cut and unambiguous. Very frequently, however, the findings are incomplete or inconclusive, so that it is necessary to make inferences with extreme caution. Careful application of the methods of statistical inference will then be necessary in order to guard against unwarranted generalizations (Chapter XIII).

(9) **Presentation of results in tables and diagrams.** Before writing the report most workers will find it desirable to prepare

rough sketches of the tables and diagrams to be used in the study. It is often convenient to cut these out and to pin them into the text as it is written (Chapter III).

(10) **Writing the report.** A satisfactory report will usually parallel in a general way the steps outlined above. It will contain a statement of the problem and its setting in the larger field; a description of the group studied; an account of the materials and methods employed; the results, inferences, and conclusions of the study; and a summary of the results obtained.

With this general plan in mind we may next turn to a detailed account of the various statistical methods.

CHAPTER II

COLLECTION AND CLASSIFICATION OF DATA

1. Primary and Secondary Data

The raw material employed in statistical studies consists in measurements or estimates known as *data,* which are *numerical statements of facts* in any department of inquiry, such as astronomy, economics, biology, psychology, and education. In the last field examples are furnished by the scores of pupils on standardized tests, physical measurements of children, salaries of teachers, attendance records, etc.

Data from whatever source may be described as *primary* or *secondary.* These terms are used in statistical method in much the same way as in historical research. In the latter field a fact taken from an ordinary text is considered as secondary material because it is removed at least one step from the original record. If the information were secured first-hand from documentary sources such as laws, original proceedings, letters, etc., it would be considered as primary historical data.

In the case of statistical method, primary data may be described as those secured from questionnaires, measurements, or estimates before the material has been combined or treated in any way so as to obscure the units or method of collection. Secondary data, on the other hand, are those which have already been collected and tabulated in some form available for use. They are usually removed one or more steps from the form of the original record, and hence comparison with similar material is of doubtful significance.

If the problem were to determine the academic training of teachers beyond four years of high school, primary records might consist of returns from a large sampling of individual

teachers' replies to a question blank. Secondary data for such a problem could be secured from the reports of state superintendents. The latter type of material would be relatively easy to obtain, but would be open to the objection that the types of teachers, units of tabulation, and other factors might not be comparable in the various reports.

Primary data are of course much to be preferred to secondary material. In case the study is of wide scope, however, and requires an elaborate plan for securing the facts, the work of collection will usually be too much for a single individual. Such studies are often subsidized by grants from public and private funds so that a staff of trained workers may gather the material. Assistance of this sort is particularly necessary in the field of school costs,* where differences in methods of accounting require personal tabulation of the data directly from the school records and invoices.

Studies which involve primary data and which may be effectively handled by a single person include experiments with apparatus or standardized tests, questionnaire investigations of limited scope, and intensive problems where the method of personal estimate or observation is required.

2. Some Examples of Secondary Source Material

The student who wishes to employ secondary material will find a large amount in government reports, school surveys, foundation publications, and funded inquiries. The Federal sources include the annual and sundry reports of the United States Bureaus of Census, Education, and Labor. Dr. Leonard P. Ayres made extended use of such material in preparing his volume, "An Index Number for State School Systems," from Bureau of Education reports.† He was able to obtain figures on

* See N. B. Henry, A Study of Public School Costs in Illinois Cities. The Macmillan Company, 1924. (This is one of the studies subsidized by the Commonwealth Fund.)

† Leonard P. Ayres, An Index Number for State School Systems. Russell Sage Foundation, 1920.

school costs and attendance for all the states running back over a period of fifty years.

Reports from state superintendent and state departments are often useful in making preliminary studies of a type reported by William R. Burgess on the academic preparation of teachers.* Dr. Burgess summarized the reports from fourteen states and found that the average teacher in 1918 had only one and one-quarter years of training beyond high school.

School surveys furnish a very valuable source of comparative data, but the variations in the methods employed for securing the financial and test data make extreme caution necessary in using such facts. The volume of the Educational Finance Inquiry on "Financial Statistics," prepared by Miss Mabel Newcomer,† is another example of a useful compilation for comparative purposes. Similar studies may be found by consulting the extensive bibliography on school costs prepared by Dr. Carter Alexander.‡

3. Units of Collection

In gathering statistical data it is usually necessary to decide in advance upon the units to be employed. For Dr. Burgess's problem, cited above, the character, "teacher training beyond high school," might have been expressed in a variety of units such as semester hours, quarters, semesters, or years. In dealing with normal-school and college training it seemed advisable to him to consider a year of training as the unit no matter where taken. The choice of such a crude unit of course makes fine comparisons of doubtful significance. Two years of training in a very poor normal school are not equivalent to two years in a first-class institution. The decision as to how rough a unit may be employed will depend largely upon the purpose

* W. R. Burgess, "The Education of Teachers in Fourteen States," *Journal of Educational Research*, March, 1921.

† Publications of the Educational Finance Inquiry, Vol. VI. The Macmillan Company, 1924.

‡ Volume IV of the Educational Finance Inquiry. The Macmillan Company, 1924.

of the study. Dr. Burgess was interested not in individual differences in teacher training but in securing an approximate index of the amount of such training in a whole state. For such purposes the unit employed was a very reasonable one.

With test data the units of collection are given by the tests themselves in terms of points, years of mental or educational age, or as functions of group variability (see Chapter VII). In recent years it has been discovered that the units in many earlier scales were expressed to a fictitious degree of accuracy. Problems in a "scaled" series were assigned values such as 3.24 under the assumption that abilities could be measured with an accuracy of one-hundredth of a "probable error" unit, as it is called. The instability of mental characters makes such precision unwarranted. Another measurement of the same person would probably differ from the previous one by a whole unit of "probable error." For most test material the simple unweighted item furnishes a unit which is sufficiently accurate for all statistical purposes, although derived scores such as mental or educational ages are often very convenient.

In the case of stable characters, such as height, greater care is needed in determining the unit of measurement to be employed. The classification of the data and the comparisons which follow will depend upon the degree of accuracy in the original material. It is usually best to make the measurements somewhat finer than the unit to be employed later in grouping the data. Thus if heights are to be classified in one-inch intervals the measurements might be made to the nearest quarter or eighth of an inch. This insures a fairly even distribution of the observations over the intervals as shown in section 8.

4. Types of Series

A *statistical series* may be described as a set of data the items of which have some common feature, or *character*. Examples of widely different characters are height, intelligence, and religious

denomination. When measurements of height, estimates of intelligence, or verbal descriptions of religious denomination are made, the resulting data may be regarded as a statistical series. It is the function of statistics to summarize, compare, and draw inferences from such series relative to some problem.

In order that appropriate methods may be used with different sorts of data, it will be desirable to have a classification of the various types of statistical series which may arise. The basis of this classification is in the representation of the character as next described.

The three characters cited above differ in certain important respects. Height and intelligence are termed *ordered characters*, because the amount or degree of either trait may vary in an orderly manner. Religious denomination, on the other hand, does not lend itself to such gradation, but must be described in verbal categories not related to one another in any orderly fashion; that is, it would be indifferent whether "Congregational" be placed before or after "Methodist" in a classification. Such characters may be called *unordered.*

A second difference between the above characters arises from the manner in which the amounts, degrees, or categories are described, or, more briefly, according to the *mode of indexing* of the character. These modes of indexing may be *numerical* or *verbal.* Thus height is said to be numerically indexed when various amounts for different individuals are expressed by the numbers arising from some scale, that is, by measurements. Intelligence may be indexed numerically or verbally. In the former case mental ages or intelligence quotients could be employed, while, for the latter mode of indexing, verbal descriptions, as "inferior," "medium," and "high," might be used. The limits of these categories may or may not be stated in numerical terms according to the manner in which the data are gathered. As already noted, the three categories of intelligence can be arranged in an orderly fashion, but in the case of religious denomination the verbal categories could be placed in any order.

A third distinction may be made between *continuous* and *discontinuous* characters. For the former the unit of measurement may be made as fine as we please, while determinations for the latter type must always be given in integers (whole numbers). Thus height is a continuous character, because gradations of height vary continuously; but size of class would be regarded as a discontinuous character, since fractional class sizes cannot occur.

The preceding analysis of characters may now be used to determine a classification of statistical series, the various types depending upon the ordering and indexing of the characters involved. Thus when the character is ordered and numerically indexed the resulting series will be called *quantitative*; when the character is ordered and verbally indexed the series will be termed *qualitative*; whereas for an unordered and verbally indexed character the series will be designated as *unordered.*

It should be noted that the basis of the present classification is in the ordering and indexing of the character and not in the nature of the trait itself. Thus both speed and quality of handwriting may furnish either quantitative or qualitative series because both may be either measured or verbally described. An objection may be made to the use of the term "qualitative" to describe series where the trait itself is not a quality in the ordinary sense. This term, however, seems a convenient one for verbal and ordered characters, and no confusion need arise if it be understood that "qualitative" is merely a label for such series.

The table on page 15 gives a brief summary of the above types of series.

5. Methods of Collecting Data

If secondary data are employed in a statistical study, it is only necessary to assemble the material from the records in some form convenient for use. Primary data may also be collected by *tabulation* of existing records, but they are usually gathered by *enumeration* or *counting* in such problems as school

TABLE 1. CLASSIFICATION OF SERIES *

ORDERING AND INDEXING OF THE CHARACTER	RESULTING STATISTICAL SERIES	EXAMPLE
Ordered		
Indexed numerically	*Quantitative*	
	Continuous	Test scores
	Discontinuous	Size of classes
Indexed verbally	*Qualitative*	Estimates of intelligence
Unordered		
Indexed verbally (possibly numerically)	*Unordered*	Classification of religion, race, occupation, etc.

census, by *estimation* as in experimental work and the appraisal of teaching efficiency, by *measurement* with physical and mental scales, or by *questionnaires* with inquiries of broad scope.

The particular method of collection to be employed will depend upon the problem and the availability of the data. It is usually best to avoid such indirect methods for securing data as the questionnaire when it is to be filled in from printed directions. The dangers of securing incomplete, unrepresentative, and faulty information from such sources are very great.

In collecting data by enumeration, estimate, or measurement it is desirable that the work be done by trained persons and by uniform methods. For a problem in child accounting, knowl-

* In his "Introduction to Statistics" Mr. G. U. Yule distinguishes between statistics of *attributes* and statistics of *variables*. For the former the observer notes only the presence or absence of some attribute and counts the number of individuals who do or do not possess it; for the latter type, determinations of some variable are made. Examples given for statistics of attributes are the number of blind and not blind, sane and insane, or tall and short persons. Measurements of height or weight furnish the data for statistics of variables.

The twofold classification given by attributes is rather restrictive and leaves open to doubt the designation of many series which may arise. Thus if another group "medium" be added to "tall" and "short," we are at a loss to know whether the resulting series is to be classified under attributes or variables. Again, if we consider the disabilities "blindness," "deafness," and "insanity" the same question arises. It appears more satisfactory, therefore, to consider the above series given by height as qualitative, since this character is ordered and verbally indexed, and to designate the disability series as unordered on account of the indifferent arrangement of the three classes.

edge of the terms and methods in this field would be necessary before comparable data could be collected from various sources. In the appraisal of methods of classroom instruction a uniform system and a careful technique of observation would be required, while for problems involving the use of standardized tests the service of trained workers in the administration of such scales is usually needed. *No elaborate statistical treatment can correct the faults of poor original data, and anything that can be done therefore to improve the reliability and accuracy of the material is time well spent.*

6. Sampling

By the sampling process is meant the use of a *sample* or portion of a larger *universe* of material taken for the purpose of drawing conclusions as to the whole. Thus if age norms are to be prepared for a certain test it is clearly impossible to examine all children of the required ages. It is therefore necessary to base the averages upon representative samples taken from the larger universe or *population.* If the samples are fairly large and properly chosen, the results will not only be very close to those which would have been obtained from the whole population, but it is also possible to predict from the sample the range within which the true value will very probably lie (see Chapter XIII). This makes it possible for the statistician to generalize beyond his actual data, and to express the so-called "reliability" of his result in terms of mathematical probability.

The principle behind the sampling process is that a fairly large number of items chosen at random from a large group or population is very likely to have the characteristics of the whole population. This may be called the *Law of Statistical Regularity for Large Numbers.*

A simple illustration of this law is furnished by an experiment to determine the percentage of Ford cars appearing on a southside boulevard in Chicago. The results found were typical of

what might be considered a universe of Fords frequenting that part of the city. The method adopted was to count 100 passing cars and note the number of Fords in such a sample with the following record:

TABLE 2. DATA ON FORD EXPERIMENT

PERCENTAGE OF FORDS OBSERVED IN ANY ONE EVENING'S SAMPLE OF 100 CARS	NUMBER OF SAMPLES OF 100 CARS EACH FOR A GIVEN PERCENTAGE OF FORDS
28	1
27	–
26	–
25	2
24	–
23	2
22	4
21	5
20	3
19	2
18	2
17	2
	Total . . . 23

The average and most frequent percentage of Fords was 21, with a maximum variation in the other samples of only 7 per cent. Any one of the twenty-three samples then gives a fairly good indication of the required percentage. It must be kept in mind, however, that the above results were for only one section of Chicago during a certain time of the day and for only one season of the year. Three samples taken on the north side two months later gave a percentage of only eight.

Another example illustrating the law of regularity in sampling is furnished by Mr. Ben Wood.* Cards for 6468 boys were filled with information regarding guardianship, number of children in the home, and other similar data. By putting the cards in alphabetical order and selecting every fourth one Mr. Wood was able to secure a sample the characteristics of which were in remarkably close agreement with those for the whole group.

* Ben Wood, "The Reliability of Prediction of Proportions on the Basis of Random Sampling," *Journal of Educational Research*, December, 1921.

A modified portion of his tables shows that a quarter of the cards, chosen as they were, was an adequate sample for comparative purposes.

TABLE 3. PER CENT OF BOYS LIVING UNDER VARIOUS HOME CONDITIONS

ITEM	PORTION OF 6468 CARDS USED		
	One Fourth	Three Fourths	All
I. Guardian			
Father	83.4	82.4	82.4
Mother	13.3	14.1	13.9
Uncle	0.6	0.6	0.6
Aunt	0.4	0.2	0.2
Stepfather	0.7	0.9	0.9
Stepmother	0.2	0.1	0.2
II. Number of children in family			
One	6.0	6.3	6.3
Two	11.3	11.8	11.7
Three	14.8	13.7	13.9
Four	13.6	14.4	14.2
Five	14.3	14.6	14.5
Six	11.9	12.6	12.4
Seven	9.8	10.5	10.3

In securing a *random sample* the principle to be kept in mind is that every individual in the group should have the same (or nearly the same) chance of being included in the sample. This is accomplished in several ways. One plan is to mix the data very thoroughly and then take a limited portion of them. This procedure is exemplified in the shuffling and dealing in ordinary card-playing. The purpose of the mixing or shuffling is to produce what is called a random distribution, a portion of which furnishes the random sample. Such distributions are assumed to be already existent in many problems, such as that of the motor cars, where the arrangement of a given one hundred cars was affected by many chance factors. The same assumption is made in measuring rainfall. Although the drops fall unevenly they tend to moisten given areas equally in the long run, and hence a gauge of a certain area furnishes a random sample. It is, of course, true that for large cities such as Chicago, samples

from different parts of the city need to be taken in order to get an adequate measure of the rainfall for the whole city. On numerous occasions it has rained in one part of the city and not in others at the same time.

Good results may often be secured by taking the items at regular intervals after the material has been arranged in some order. In Mr. Wood's experiment every fourth card was selected under alphabetical arrangement. This plan is usually satisfactory unless there is some reason to expect a relationship between the character studied and alphabetical order. Thus, in a study involving pupil recitation there might be a tendency on the part of some teachers to call more frequently on pupils whose names begin with the earlier letters of the alphabet.

If the population sampled contains a number of types, a purely random sample of the whole is probably not best because some of the types may be omitted or not fairly represented. For such problems sub-samples proportional in size to the numbers in the various types should be selected. For example, in a study of high-school pupils samples from each of the four years might be chosen and combined, the size of the samples being taken proportional to the relative numbers of pupils in the four high-school classes.

The size of the sample will depend upon the degree of accuracy required in the result, the precision varying as the square root of the number of cases. As indicated in the first chapter, forty to sixty cases are as few as can be expected to yield good results in experimental work. When only fifteen or twenty are used, the application of the usual laws of sampling becomes very doubtful.

7. Arrangement of the Original Data

The form of the permanent and working records will depend upon the number of data employed. The *master sheet*, as shown in Exercise 1 at the end of this chapter, is advisable for samples of fifty to one hundred cases. With a large amount of material,

however, a uniform blank card or page is usually required. Sir Francis Galton kept a record of his data in large bound volumes with a page for each person examined, the age, profession, nationality, and the results of various mental and physical tests being set down in the appropriate spaces.

If the series is short, the tallying and distributions may be made directly from the master sheet by running down the page and checking off the items. This method, however, makes it necessary to go over the whole list to catch a single error and is rather awkward for the preparation of correlation tables because the order and accuracy of entry produce a strain on the attention.

Case No.	C. A. 16.4	M. A. 16.4	I. Q. 100	5.31
38	.142	51	13.2	65.5
252.	.974	13.2	37	18.91
47.31	14.67	68.1	.061	1.121
4.2	72.4	39.12	9.8	16.5
37.41	.22	2.65	207.	.0111

FIG. 1. Data Card

The preparation of small tickets for a working record will overcome most of the above difficulties. These cards should be fairly thin, of uniform size, and have a corner-cut off to facilitate the separation into piles during the sorting. The data are set down from the permanent record in the form of numbers with a definite spatial arrangement on the card. In order to identify the tickets with the permanent record the case number should appear on a corner of the card. This will make it possible to prepare a duplicate in case a card is lost. The size of the card will depend upon the number of items entered, but it should be as small as can be handled conveniently. A key for the items entered will of course be required for a sample card such as the one shown in Fig. 1. In sorting, the characters are easily identified by their position on the card.

In case the work of tabulation and sorting is to be done by mechanical devices such as the Hollerith Machine, the data card will be a convenient record when punching the information for the tabulating card. The holes in this card (Fig. 2) make it possible to sort very rapidly by electrical contact.

Kind of School	Month	School Number	Class Number	Teacher Number	Days Present	Days Absent	Enrollment: Oakland Boys	Enrollment: Oakland Girls	Enrollment: State Grade	Enrollment: State Boys	Enrollment: State Girls	Enrollment: End of Month Grade	Enrollment: End of Month Sect. A	Enrollment: End of Month Grade	Enrollment: End of Month Sect. B	Enrollment: End of Month Total	Vacant Seats	Absent on Account of Illness		Spe'c. Pro'tns	Demotions
Da. El.	0	0 0	0	0 0 0	0 0 0 0	0 0 0	0	0	0	0	0	0	0	0	0 0	0 0	0	0 0	0 0	0	0
Ev. El. 1	1	1	1 1	1 1 1	1 1 1 1 1	1 1	1 1	1 1	1	1 1	1 1	1	1 1	1	1 1	1 1	1	1 1	1 1	1	1
Da. Hi. 2	2	2 2	2	2 2 2	2 2 2 2 2	2 2	2 2	2 2	2	2 2	2 2	2	2 2	2	2 2	2 2	2	2 2	2 2	2	2
Ev. Hi. 3	3	3	3 3	3 3 3	3 3 3 3	3 3	3 3	3 3	3	3 3	3 3	3	3 3	3	3	3 3	3	3 3	3 3	3	3
Dept'l 4	4	4 4	4 4	4 4 4	4 4 4 4	4 4	4 4	4 4	4	4 4	4 4	4	4 4	4	4	4 4	4	4 4	4 4	4	4
Kgt'n 5	5	5 5	5 5	5	5 5 5 5 5	5 5	5 5	5 5	5	5 5	5 5	5	5 5	5	5 5	5 5	5	5 5	5 5	5	5
6	6	6 6	6 6	6 6 6	6 6 6 6 6	6 6	6 6	6 6	6	6 6	6 6	6	6 6		6 6	6 6	6	6 6	6 6	6	6
Year 7	7	7 7	7 7	7 7 7	7 7 7 7 7	7 7	7 7	7 7	7	7 7	7 7	7	7 7	7	7 7	7 7	7	7 7	7 7	7	7
8	8	8 8	8 8	8 8 8	8 8 8 8 8	8 8	8 8	8 8	8 Kgn.	8 8	8 8	8 Spol.	8 8	8	8 8 Ung.	8 8	8	8 8	8 8	8	8
9	9	9 9	9 9	9 9	9 9 9 9 9	9 9	9 9	9 9	9	9 9	9 9	9	9 9	9	9 9	9 9	9	9 9	9 9	9	9

FIG. 2. Hollerith Tabulating Card used in Oakland, California, School System

8. The Simple Frequency Distribution

In dealing with a large body of data it is necessary to classify the material in some compact and orderly form before it can be effectively analyzed. The frequency distribution is the most convenient arrangement for the material, because it reveals some of the most important properties at a glance and makes all of the calculations very much easier than would be possible with the ungrouped items. A simple frequency distribution consists of a series of *classes* of the character and a set of corresponding *frequencies*. In the case of a quantitative series the scale is usually divided into a number of *classes* of equal width, for example, 54.5 to 59.5, 59.5 to 64.5, 64.5 to 69.5, etc. The number of items or measures (called the frequency) occurring in each interval is then determined by tallying. For qualitative or unordered series the classes are indicated verbally and the frequencies tabulated as in the first case.

The ancient method of tallying is to record the frequencies by strokes until four have been made and then to make a cross stroke. This makes it easy to count the marks. For example:

Class	Tally	Frequency
64.5–69.5	𝍸 /	6
59.5–64.5	𝍸 𝍸 /	11
54.5–59.5	𝍸 //	7

The tally marks, of course, should not appear in the final distribution.

The steps in making a frequency distribution for a quantitative series consist in (1) noting the *range* of the data, that is, the distance between smallest and largest items; (2) deciding upon the number of classes into which the material is to be grouped; (3) determining the numerical limits of the classes; and (4) tallying the frequencies in the appropriate classes. Steps (2) and (3) are important because all of the subsequent calculations will be affected by the width and limits of the

classes. The data, when grouped, are considered to be either concentrated at the midpoints of the intervals or spread evenly over them. The calculations from grouped data will then not agree exactly with those from ungrouped series unless the width of the classes is equal to the collection unit. If the grouping is this fine, however, the classes may be so numerous that the advantage of employing a distribution is lost and the frequencies are likely to present a very irregular appearance, not typical of the continuous gradation expected from ordered characters. It is therefore better to use a wider interval smoothing out the accidental irregularities, probably due to sampling, and making the subsequent calculation easier although slightly less accurate. When there are from fifteen to twenty-five classes with material consisting of one hundred or more items, the error due to grouping is very slight, and even this may be adjusted by certain corrections (see Chapter XVI, section 8).

The choice of class limits depends upon the accuracy of the original data. If the measurements are very fine and the classes fairly broad, the limits of the classes may be expressed in the form 55–59.99, 60–64.99, 65–69.99, etc. This method makes it possible to assign measurements very definitely to the appropriate classes, since all items equal to the lower limit and up to but not including the upper limit are located in a given class. One difficulty with this designation, however, is that confusion sometimes arises regarding the numerical value of the upper-class limits in calculation. Students may take these to be actually 59.99, 64.99, etc. An alternative plan is to write the classes in the form $55–60^-$, $60–65^-$, $65–70^-$, etc., with the understanding that 60^- is equal to 60 for purposes of calculation, but means just less than 60 in the tabulation.

A more important objection to the above method arises when the measurements are not very fine. If we assume that the items are given correct to the nearest integer, an even distribution of the observations over a class interval would be represented as shown in Fig. 3, p. 24.

The *class values*, or midpoints of the intervals used in subsequent calculations, will thus be 57.5, 62.5, etc. These values, however, are not representative of the items in the respective classes. In the first interval, for example, there are three items below 57.5 and only two above, spacing the observations unequally about this class value.

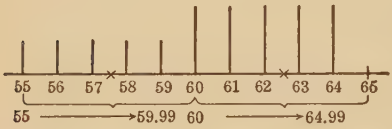


FIG. 3. Illustrating the incorrect class limits which appear when the items are measured to the nearest integer

In order to adjust for this difficulty, the class limits should be set as in the diagram shown in Fig. 4 by moving one half of the collection unit down on the scale. The location of any frequency is as uniquely determined as before, and the class values 57, 62, etc. are more truly representative of the items in the intervals 54.5–59.5, 59.5–64.5, etc.

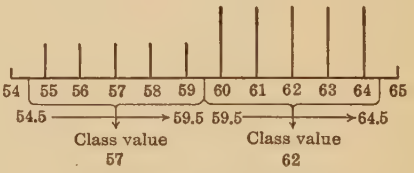


FIG. 4. Illustrating the correct method of stating class limits when the items are given to the nearest integer

If the measurements are so fine that practically the whole interval 55–59.99 could be filled with observations, then 57.5 would be an approximately correct class value and the

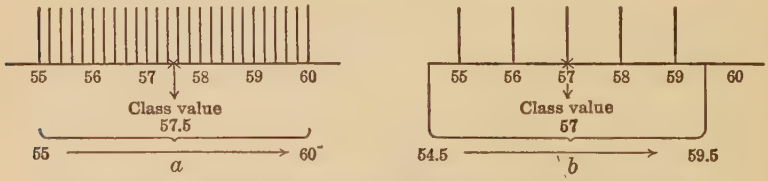


FIG. 5. Illustrating class values and limits with measurements

a, very fine; *b*, to nearest integer

limits 55–60⁻ satisfactory. If, on the other hand, the measurements are rather coarse, the size of the collection unit needs to be taken into account by moving the interval back one half

the amount of this unit. If this is not done, all class values (and the resulting average for the whole series) will be one half of the collection unit too large.

The following frequency distribution has been made from the Otis Test Scores appearing in Exercise 1 at the end of this chapter. Inasmuch as the scores are given to the nearest integer (point), the classes will run from 79.5–89.5, 89.5–99.5, etc.

TABLE 4. FREQUENCY DISTRIBUTION FOR OTIS TEST SCORES

CLASS	TALLY	FREQUENCY
179.5–189.5	/	1
169.5–179.5	/	1
159.5–169.5	////	4
149.5–159.5	~~////~~ ~~////~~ /	11
139.5–149.5	~~////~~ ////	9
129.5–139.5	~~////~~ ~~////~~ /	11
119.5–129.5	~~////~~	5
109.5–119.5	////	4
99.5–109.5	//	2
89.5–99.5	/	1
79.5–89.5	/	1
		Total 50

It might be argued that a person receiving a score of 80 could not have done less than the amount required to receive such a score, and that he very probably did a little more, so that his truer score for that performance should be 80.5 instead of 80. This reasoning would lead to class intervals, 80–90⁻, 90–100⁻, etc., but would be contrary to the usual practice of taking scores at their face value. In the present discussion, therefore, we shall assume that scores are correct to the nearest integer.

It will be noted that only eleven classes were used in this distribution because of the small number of cases involved.

9. THE CLASSIFIER

In dealing with small samples it is frequently desirable to rank the items and to prepare short frequency distributions. For such purposes a device known as the *classifier* will be found

very convenient. It consists of tabular arrays of small cells identified by units' digits on one axis, and by tens' digits on the other. The location of any item is then readily indicated by a tally mark in the appropriate cell. The accompanying classifier has been made for the Otis scores given in Exercise 1, p. 29.

TABLE 5. CLASSIFIER FOR OTIS TEST SCORES *

TENS	UNITS										TOTALS
	0	1	2	3	4	5	6	7	8	9	
18								/ 1			1
17		/ 2									1
16		/ 6					/ 5	/ 4		/ 3	4
15	// 16.5	// 14.5	/// 12	/ 10				/ 9	/ 8	/ 7	11
14	/ Md. 26 X	/ 25	/ 24		/ 23	// 21.5	// 19.5			/ 18	9
13		// 36.5	/ 35	// 33.5	/ 32	/ 31	/ 30	/// 28			11
12					/ 42	/ 41	/ 40		// 38.5		5
11		/ 46			// 44.5				/ 43		4
10		/ 48					/ 47				2
9						/ 49					1
8								/ 50			1
Totals .	3	9	5	3	5	5	6	7	4	3	50

* This useful device was first brought to the attention of the writer by Dr. Leonard P. Ayres in a series of lectures given at The University of Chicago in 1920. It is recommended for use when dealing with fifty to one hundred cases.

The scores are entered in the classifier just as they come from the master sheet. Thus the first Otis score from the list is 171. Looking down the left-hand margin for the tens' digit 17, and moving across under the units' digit 1, locates the tally in the proper cell. To check the work the tallying may be repeated by making a small dot over each tally stroke.

It will be noted that the material has been arranged in classes ten units in width and that the distribution in the totals on the right is the same as that found in section 8. The distribution of the totals at the bottom of the classifier is a random arrangement, the number of scores ending in 0, 1, 2, 3, etc. tending to be the same in the long run.

The numbers in the cells indicate the ranks of the various items, and are determined after all of the tallying is completed, by counting down from the highest score. The advantage of the classifier for ranking is that if the tallying is correct, none of the scores will be omitted as would be quite likely if they were arranged in rank order by searching in the list of fifty for successively smaller items.

When a score of 152 has been reached in the ranking, three tallies will be found. Inasmuch as these have the same value it is customary to assign to each the average rank of 11, 12, and 13, which is 12. In the same way the two scores of 151 would share the next two ranks 14 and 15, each being given the average rank of 14.5.

In addition to the grouping and ranking of the data the classifier will also be found useful in determining the *median*. This average for ranked items is the middle score, or is halfway between the two middle scores, for an even number of items. In the problem above the median is 140.5 by inspection.

It will be noted that the median is here defined as the middle score. For an odd number of cases this definition offers no difficulty, but with an even number the use of the value halfway between the two middle scores is a convention supplementary to the definition.

10. Cumulative Frequency Distributions

It is often useful to have the data arranged in a cumulative rather than a simple frequency distribution. This is accomplished by tabulating all of the frequencies less than the upper limit of each class interval. For the Otis material the cumulative distribution would be as follows:

Table 6. Cumulative Frequency Distribution for the Otis Test Data

Score Less Than	Cumulative Frequency
189.5	50
179.5	49
169.5	48
159.5	44
149.5	33
139.5	24
129.5	13
119.5	8
109.5	4
99.5	2
89.5	1

The cumulative frequencies are of course easily tabulated after the simple frequency distribution has been made. Both methods of representing series will be extensively used in applying the descriptive methods of the following chapters.

EXERCISES

1. Make a classifier for the fifty Terman scores in the table on page 29 and obtain the ranks of the scores. Determine the median from these ranks. (123 or 122.7. *Ans.*)

2. Work out a scheme for ranking the Chicago scores and obtain the median. (53.25. *Ans.*)

3. Make a simple frequency distribution for the Terman scores from the classifier. The classes will be 169.5–179.5, 159.5–169.5, etc. Make a similar distribution for the Chicago scores with classes 74.75–79.75, 69.75–74.75, etc.

SCORES OF FIFTY PUPILS ON THREE INTELLIGENCE TESTS

PUPIL	TEST			PUPIL	TEST		
	Otis	Chicago	Terman		Otis	Chicago	Terman
1	171	52	117	26	133	47	101
2	169	75.5	153	27	151	53.5	137
3	128	50.5	131	28	145	56.5	119
4	141	46	105	29	152	56	124
5	106	39.5	71	30	157	66.5	170
6	146	55	130	31	144	60.5	155
7	87	34	80	32	140	60.5	119
8	114	42	101	33	111	38.5	142
9	187	70	153	34	150	63.5	140
10	133	51.5	132	35	152	65.5	122
11	151	59	136	36	137	48	115
12	131	52.5	128	37	146	54	125
13	150	63	145	38	128	44.5	87
14	118	44.5	110	39	145	57.5	120
15	142	65.5	122	40	153	50.5	117
16	166	61	152	41	149	53	135
17	158	55	157	42	114	45.5	100
18	101	39	88	43	135	40	125
19	159	57.5	156	44	131	47	120
20	126	41.5	92	45	161	61	149
21	136	65.5	115	46	95	37	87
22	137	63.5	109	47	134	50	103
23	152	75.5	151	48	124	48.5	119
24	137	45	132	49	125	43	95
25	132	61.5	130	50	167	58.5	178

4. Arrange separately the Terman and Chicago scores in the form of cumulative frequency distributions.

5. Make a frequency distribution for the following scores, using an interval of one unit: 11, 12, 12, 13, 13, 13, 14, 14, 14, 14, 15, 15, 15, 15, 15, 16, 16, 16, 16, 17, 17, 17, 18, 18, 19. Calculate the average (mean). What will the average be if the intervals are taken 11–11.99, etc., instead of 10.5–11.5, etc.? What is the error in the average by the former tabulation method? (Error is .5.)

6. Tabulate separately the scores on page 30 on spelling tests A and B for 125 pupils, using an interval of 5.

7. Retabulate the scores in Exercise 6, using an interval of 10. Which interval is better?

8. Make cumulative frequency distributions from the two spelling test distributions of Exercises 6 and 7.

SCORES OF 125 PUPILS ON TWO SPELLING TESTS OF EQUAL DIFFICULTY
(Maximum Score = 105)

TEST		TEST		TEST		TEST		TEST	
A	B	A	B	A	B	A	B	A	B
43	44	91	84	57	59	45	37	65	68
38	41	56	55	82	87	52	49	86	87
83	87	66	61	53	61	57	69	25	25
57	65	15	18	64	68	91	96	48	57
63	70	47	62	64	62	84	78	71	73
92	89	15	21	73	66	89	89	64	62
68	63	63	70	74	76	81	76	48	57
94	91	44	39	54	58	90	92	74	74
79	84	100	95	68	77	70	76	86	93
45	45	68	73	29	35	87	91	33	29
79	81	26	30	43	53	83	79	42	44
65	70	83	79	54	56	45	48	92	97
20	33	102	101	85	87	93	89	83	88
93	91	83	70	19	26	82	74	59	64
67	72	85	85	81	80	55	57	63	62
59	61	21	27	86	83	37	41	56	48
98	99	81	76	74	77	31	26	6	6
81	84	51	46	68	75	39	42	27	37
86	79	49	52	69	62	16	16	41	52
57	71	67	61	58	54	95	96	25	31
92	82	86	80	85	85	68	71	55	59
30	35	37	35	40	36	48	56	52	55
80	79	43	49	79	90	38	49	68	77
68	72	46	47	75	80	63	68	25	22
83	75	85	83	63	55	80	86	53	59

CHAPTER III

TABULAR AND GRAPHICAL PRESENTATION OF DATA

1. Purpose of Tables and Diagrams

Although the preparation of tables and diagrams will usually be the last step in working out a statistical problem, it is well to consider such work at this point because of its relative simplicity and concreteness. For many elementary studies, moreover, such as school and publicity reports, the tabulation and graphical representation of secondary material is about the only statistical method required. It is, therefore, desirable that everyone dealing with educational statistics should become acquainted as soon as possible with simple tables and graphs.

In the following discussion the word "diagram" is used to describe all sorts of graphs, charts, plots, or maps used for the display or comparison of data.

Tables and diagrams have a twofold purpose: one is to assist in the analysis of the material and simplify the calculations by representing the data in concise and orderly fashion, while the other is to summarize and make clear the findings of a study. Thus the chief reason for arranging material in a frequency table is to facilitate analysis and calculation. The important characteristics of the series may then be readily determined and the required calculations made more easily than from the ungrouped data. On the graphical side a method of calculation has been developed known as *nomography*. By means of curves drawn to suitable scales a great many statistical calculations may be made very quickly. In many cases, however, the construction of the nomograph is very laborious and the desired calculations will not be given to a sufficient number of significant figures. With the modern development of calculat-

ing machines and statistical tables for computation, almost every sort of calculation will be found to be easier, more rapid, and much more accurate by numerical rather than by graphical methods.

The proper use of tables to summarize numerical results is important because the success of a statistical study may depend a great deal upon the skill with which the tabular material is arranged. Good tables are usually brief and so titled as to be self-explanatory. By a suitable arrangement of headings a large amount of important information can be given in a very short space, comparison between similar items facilitated, and visualization of group relationships made possible.

Graphs or diagrams for presentation are intended to make the numerical comparisons clearer and more vivid. They are not primarily intended to summarize the statistical findings, which should appear in tabular form accompanying the diagram. If too many details are given in a chart its clarifying value is lost and a diagram that is not clear is probably not worth making at all.

2. The Construction of Tables for Presentation

While there is not universal agreement as to the terms used and the best form for a table, the following suggestions have the merit of successful usage in the publications of the Russell Sage Foundation.

DEFINITIONS OF THE PARTS OF A STATISTICAL TABLE

1. *A statistical table* is a quantitative presentation of facts by means of numbers arranged in a column or columns and distributed according to one or more groupings of the subject matter.

2. *A table title* is a statement appearing at the head of a statistical table, showing the subject with which the table deals.

3. *A column* in a statistical table is the series of numbers, generally relating to the same unit, arranged vertically in the table.

4. *A line* in a statistical table is a series of numbers arranged in a horizontal row in the table.

5. *The body* of a table is the aggregate of the columns and the lines.

6. *A column heading* is a word or group of words at the head of a column of numbers in a table, showing the unit dealt with and the relation of the column to the classification followed.

7. *A brace heading*, or *box heading*, is a word or group of words appearing above two or more columns of numbers in a table, which it has the effect of uniting as with a brace, and to each of which it bears the same relation. In connection with the column headings, the brace heading shows the unit dealt with and the relation of each column to the plan of classification followed.

8. *A line title* is a word or group of words at the left of a horizontal line or row of figures in a table, showing the relation of the line to the plan of classification followed.

9. *A total* is a statement of the aggregate of two or more numbers appearing in a column or line.

10. *A grand total* is a statement of the aggregate of several totals.

TABLE 7. EXPENDITURE PER INHABITANT FOR OPERATION AND MAINTENANCE OF SCHOOLS IN CLEVELAND, AND IN SEVENTEEN OTHER CITIES OF FROM 250,000 TO 750,000 INHABITANTS, 1914

CITY	ESTIMATED POPULATION IN 1914 (IN THOUSANDS)	EXPENDITURE FOR OPERATION AND MAINTENANCE		RANK IN EXPENDITURE PER INHABITANT
		Total (in Thousands)	Per Inhabitant	
Baltimore	580	$1955	$3.37	17
Boston	734	5517	7.52	2
Buffalo	454	2450	5.40	12
Cleveland	639	3570	5.59	8.5
Detroit	538	2553	4.75	14
Indianapolis	259	1410	5.44	11
Jersey City	294	1421	4.83	13
Kansas City	282	1761	6.24	7
Los Angeles	439	3707	8.44	1
Milwaukee	417	1795	4.30	15
Minneapolis	343	2148	6.26	6
Newark	389	2699	6.94	3
New Orleans	361	1098	3.04	18
Pittsburgh	565	3602	6.38	5
San Francisco	449	1879	4.18	16
Seattle	313	1751	5.59	8.5
St. Louis	735	4085	5.56	10
Washington	353	2392	6.78	4
Average	—	—	$5.59	—

The model table on page 33 illustrates all of the terms used with the exception of the totals, which were not necessary. It will be noted that the basic information upon which the comparisons are made is given in the table so that it could be verified by the reader in case of doubt. The style notes used in the construction of the table are given in the following list:

STYLE NOTES FOR MAKING TABLES

I. Arrangement of Data

1. A short table is clearer and more forceful than a long one.
2. Original data should be presented in full.
3. It is easier to compare numbers arranged one above the other than numbers placed side by side. Tables should be arranged so that, as far as possible, numbers to be compared are in the same column.
4. Items listed in a table should usually be arranged in descending or in ascending order of their rank in the trait in which they are being compared.

II. Titles and Headings

1. The titles should always go above a table since a table is essentially a list.
2. Titles and headings should be so worded and the table so arranged that the result will be a complete whole, independent of the accompanying text.
3. Table titles should place emphasis upon the fact or facts which the table is intended to show. This can be accomplished by placing the important facts at the beginning of the title.
4. Words like "table showing," "number of," and "distribution of" should be omitted wherever the meaning of the title is clear without them.

III. Punctuation

1. Table titles should be capitalized throughout.
2. In column headings and in line titles, except in the case of proper names, only the initial letter of the first word should be capitalized (also good form to capitalize first letter of each word).
3. Do not end a title with a period. If the title consists of two sentences, put a period after the first sentence.

4. Do not use periods in column headings or in line titles except for abbreviations and to separate sentences as above. Avoid abbreviations when possible.

5. Do not use periods in the body of a table except to separate dollars from cents or units from tenths.

6. Where one line of items is to be compared with those in the rest of the table this line may be in heavier type, so that it may be more readily seen.

IV. Symbols

1. Ditto marks should not be used either in the body of a table or in its headings and titles.

2. Where sums of money are stated in columns, the dollar sign should be placed before the first item in the list and before the total or average.

3. Footnotes to the table should be indicated by letters and not by figures (also good form to use symbols such as *, §, etc.).

4. Where data are not available do not fill in the space in the table with 0's. Reserve 0 for the definite information that it gives, that is, nothing; use - - - - - - - or · · · · · · to show that no figures are at hand.

5. A row of dots or dashes on the *lower* part of the line may be used in the first column to guide the eye from each item to its corresponding figure. These dots should not extend beyond the first vertical rule.

V. "Total" and "Per Cent"

1. "Total" should always be written in the singular.

2. "Per cent" should be written in two words, with no period.

VI. Ruling

1. There should be a double rule at the top of the table.

2. A single horizontal rule should separate column headings from the body of the table.

3. At the bottom of the table there should be a double horizontal rule.

4. Totals and averages should be separated from the numbers of which they are the aggregates, by single heavy rulings (single light ruling is also good form).

5. There should be vertical rules between the line titles and the figures, and between each two columns of figures.

6. Tables should not be closed in at the sides by vertical rules.

7. Each column heading should be boxed in except at the two outer sides.

8. These rules may be summarized as follows: There are three kinds of lines used in ruling a table: double lines at the top and bottom of the table; single lines between column headings and figures, and between columns; and heavy lines before totals and averages.

VII. Spacing

1. In long tables it is well to leave a double space after each five or ten lines of figures, to facilitate the reading.

2. Numbers should be placed in the middle of the column with corresponding units directly under each other.

3. Column and Bar Diagrams

For a full account of the great variety of diagrams which may be used the reader is referred to such texts as Williams's, listed with the selected texts in the bibliography. The discussion here will be confined to a few simple types which serve most of the purposes in an ordinary statistical study and which can be made without much training or great outlay of drawing materials. If elaborate figures are required it is probably better to have them drawn by an artist from a rough sketch rather than spend time in acquiring the skill necessary to use a drawing board and instruments. For the great majority of articles, books, and theses, however, only the simplest types of diagrams are necessary, and these may be made on ruled paper in black ink with very little practice.

The *column diagram* consists of a series of columns proportional in height to the quantities represented. A scale usually appears at the left and a legend either on the background near the columns or below as in Fig. 6. In this figure two varying quantities are shown very effectively on the same chart, the hatched portion representing the undesirable condition. Such a diagram may be made with india ink on ruled graph paper,

blue lines being preferred, because these will be invisible if the chart is photographed.

Fig. 7 is another ingenious variation of the column diagram. Each block represents a school identified by number so that it is possible to compare any school with another or with the whole group. This type of diagram can be effectively used to represent a group of test scores in such a way that each pupil can recognize his score by number without revealing this fact to the rest of the class.

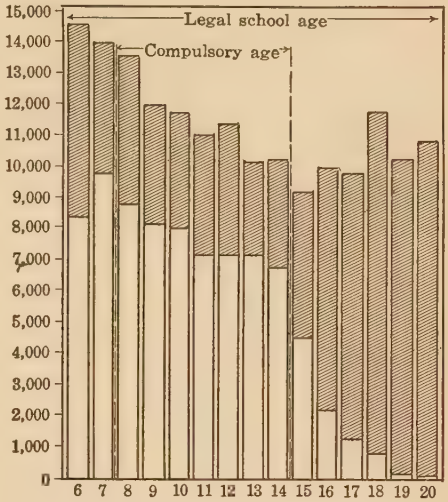


FIG. 6. Showing the holding power of the schools

The columns represent the children enumerated by the school census as of each age from six through twenty. Portion in outline represents children in public schools. Portion in black represents those not in public schools. (*Cleveland Education Survey Report*, 1916)

In case the columns are used to represent the frequencies of the various classes along the horizontal scale the resulting diagram is known as a *histogram*. The columns are then proportional in height and area to the frequencies, and this property makes the histogram an excellent representation of a frequency distribution. The histogram for the Otis scores in Table 4 is given in Fig. 8. It will be noted that the horizontal scale is given in even integers and the column moved slightly to the left so as to have the intervals 79.5–89.5 etc.

An alternative representation of the frequency distribution is given by the *frequency polygon*. This consists of lines connecting the frequencies taken at the midpoints of the class intervals. In Fig. 9 a histogram and frequency polygon are plotted

on the same background for comparison. It will be noted that the area under the histogram between I.Q.'s from 90 to 100 is

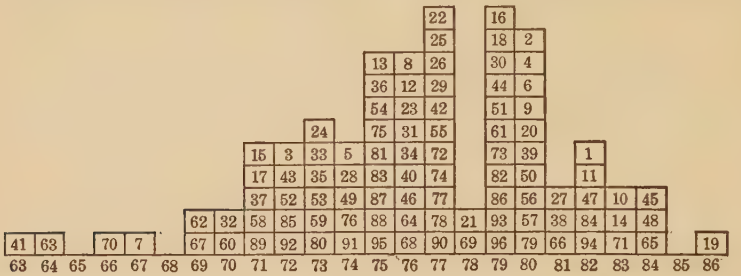


FIG. 7. Average scores made in spelling by ninety-six elementary schools*

The figures below the diagram show the percentages, and the ones in the diagram show the number of the schools

exactly proportional to the observed frequency over that range. The area under the polygon over this same range is somewhat in defect, however, and similar discrepancies occur for the other intervals unless three points on the polygon happen to be on a line. For these reasons the histogram is a better representation of the frequency distribution when a curve is to be fitted to the data. In case a rough diagram is required to show the overlapping of several distributions, the polygon is probably better, but for most other representations the histogram is preferable.

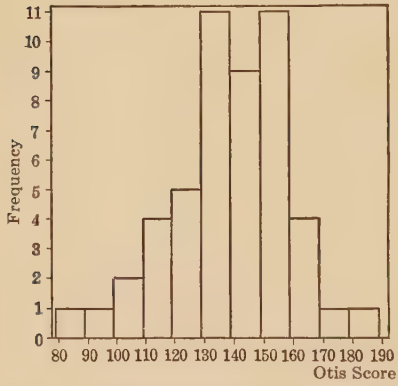


FIG. 8. Histogram of Otis Test Scores

In *bar diagrams* the varying quantities are represented by horizontal bars as in Fig. 10. The chief reason for preferring a bar to a column arrangement is one of convenience. If the line

* From C. H. Judd, "Measuring the Work of the Public Schools," *Cleveland Education Survey Report*, 1916, p. 84.

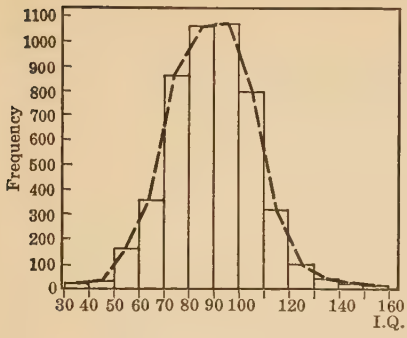


FIG. 9. Histogram and frequency polygon for the intelligence quotients of Table 20, Chapter VII

titles are large as in the accompanying figure, the use of columns would be awkward. For a fairly large number of items the bar diagram will also be found to be more effective.

Quantities which exhibit variation in one dimension should be represented by column or bar diagrams which are themselves one-dimensional. The use of three-dimensional diagrams, such as a row of persons of varying size to show increase in population, may be very misleading

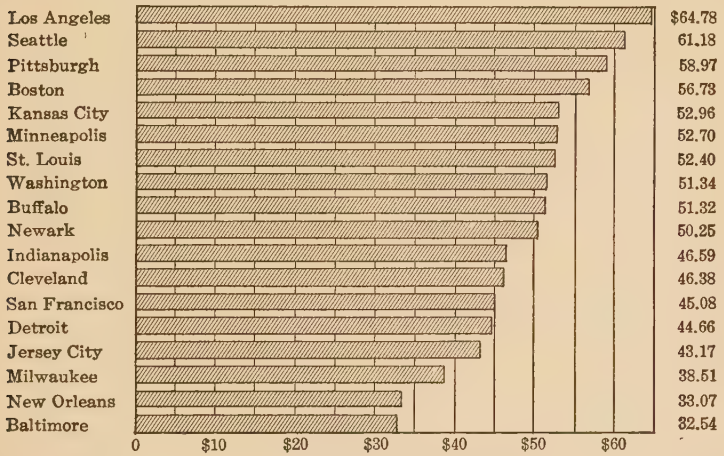


FIG. 10. Expenditure per child in average daily attendance for operation and maintenance of public schools, for Cleveland and for seventeen other cities *

because there is doubt as to whether the height, area, or volume of the figures is proportional to the change in population.

* From Earl Clarke, "Financing the Public Schools," *Cleveland Education Survey Report*, 1916, p. 37.

4. Coördinates

In order to remind the reader of his coördinate geometry, which will be very much needed in the work which is to follow, the next few paragraphs will be devoted to a summary of the elements of that subject.

If two straight lines OX and OY intersect in a plane, it is possible to describe the location of any point P in the plane with respect to the point of intersection O. For most representations it is convenient to have the lines intersect at right angles. The horizontal line OX is known as the x-axis, or axis of abscissas, while the vertical line OY is called the y-axis, or axis of ordinates. The distances of the point P from the two axes are known as coördinates of the point. Thus in Fig. 11 the *abscissa* of the point P is OM, or four units, while its *ordinate* is ON, or three units. These two coördinates will locate uniquely the position of any point P with respect to the origin O.

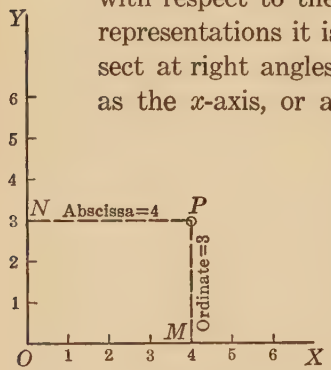


Fig. 11. Illustrating ordinate and abscissa

It will be noted that in Fig. 11 only positive quantities can be represented. In case negative numbers occur, the coördinate system may be extended as shown in Fig. 12. The plane is thus divided into four *quadrants* numbered in counterclockwise direction about O. The coördinates of a point in the second and fourth quadrants are opposite in sign, while those for a point in the third quadrant are both negative. The coördinates of the four points in the diagram are as follows:

Point	Abscissa, X	Ordinate, Y
P_1	$+4$	$+2$
P_2	-5	$+3$
P_3	-3	-2
P_4	$+7$	-1

In plotting mathematical relationships it is usually necessary to employ the more extended scheme with four quadrants, but in dealing with statistical data which are usually positive, the first quadrant will suffice.

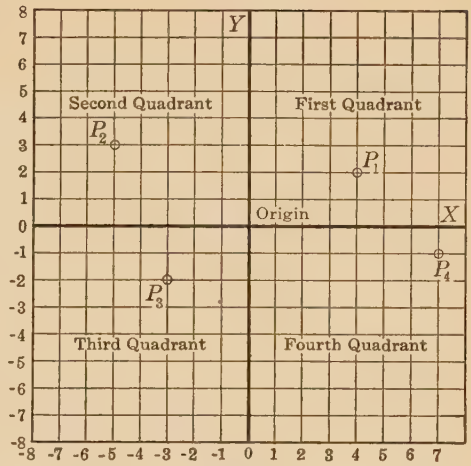


FIG. 12. Illustrating plotting in four quadrants

The following table gives the lung capacity in cubic inches of 521 boys in the laboratory schools of The University of Chicago. The ages of the boys ranged from five to nineteen years, the measurements being made within a few days of each birthday.

TABLE 8. LUNG-CAPACITY DATA FROM THE LABORATORY SCHOOLS

AGE	AVERAGE LUNG CAPACITY
5	76
6	73
7	88
8	95
9	106
10	122
11	129
12	148
13	165
14	184
15	211
16	230
17	252
18	264
19	287

These data have been plotted in Fig. 13 and the points connected in the form of a polygon. The trend appears fairly straight with

the exception of a general dip during the years of adolescence. This dip has been verified by other material.*

Such a plot as that shown below is of value in analyzing the data and in giving to the reader a clear idea of the relation between the variables involved. We shall turn next to the general consideration of such functional relationships.

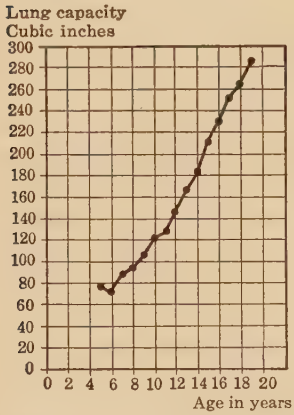


FIG. 13. A plot of lung-capacity data

5. FUNCTIONAL RELATIONSHIPS

When two variables are so related that the value of the first variable depends upon the value of the second variable, then the first variable is said to be a function of the second. The area of a square, for example, is a function of the length of the side; that is, area equals $(\text{side})^2$, or $y = x^2$. Here the relationship is exact, all true squares conforming precisely to the law. Such functional relationships may be called mathematical, and are generally written in the form $y = f(x)$.

The second variable, to which values may be assigned at pleasure, is called the *independent variable*, or *argument*; and the first variable, whose values are determined as soon as values of the argument are assigned, is called the *dependent variable*, or the *function*. In the above example the side x, representing the side of the square, is the independent variable, while y, representing the area of the square, is the dependent variable.

Other examples of functions are breathing capacity, which is a function of the age of the person; the number of words typed per minute, which is a function of the hours of practice; and the score on an achievement test, which is a function of the

* Karl J. Holzinger, "On the Relation of Vital Capacity to Certain Psychical Characters," *Biometrika*, Vol. XVI, p. 139.

time spent in studying the subject tested. Such functions differ from exact mathematical functions in that they depend upon many more variables than the ones given, and the relationships indicated are only approximate. Breathing capacity, for instance, depends upon a great many factors other than age. A curve or an equation expressing the most probable breathing capacity for given ages will then furnish a basis for rough estimation rather than exact prediction. An important part of statistical method is concerned with the selection of those mathematical functions which will give the best "fit" for a given body of data (see Chapter XVI).

6. The Straight Line

One of the simplest mathematical functions is that wherein the change in y is directly proportional to the change in x, for example, $y = 3x$ or $y = \frac{1}{2}x$. The graphs of these functions will be straight lines through the origin, as shown in Fig. 14. In obtaining the coördinates of various points it is only necessary to substitute arbitrary values for the argument x, and find the corresponding values of y. While only two points are necessary to determine a straight line, one other value has been given as a check.

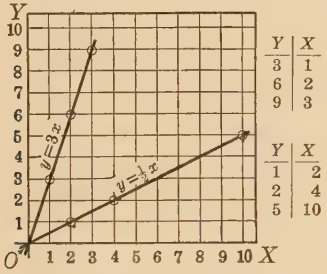


Fig. 14. Graphs of the lines $y = 3x$ and $y = \frac{1}{2}x$

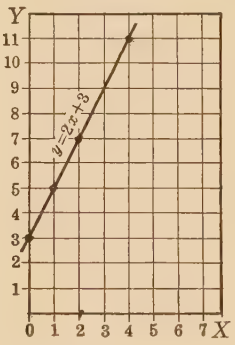


Fig. 15. Graph of $y = 2x + 3$

The general equation of a straight line may be written in the form $y = ax + b$, where b is a constant representing the distance from the origin to the point of intersection of the given line and the y-axis (*y-intercept*), and a is a constant representing the *slope* of the line (the tangent of the angle which the line makes with the x-axis). The line $y = 2x + 3$ is shown in Fig. 15.

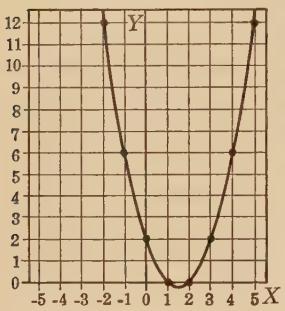


FIG. 16. Graph of $y = x^2 - 3x + 2$

7. NON-LINEAR RELATIONSHIPS

The term "curve" is employed in mathematics to designate any line, straight or curved, when located with reference to some coördinate system. It has been noted that equations of the first degree in x furnish straight lines when graphed. In case higher powers of the argument are present, some other form of curve results. One of the simplest of these is the parabola the general equation of which is $y = ax^2 + bx + c$, where the letters a, b, and c again represent constants which determine the particular curve. The parabola $y = x^2 - 3x + 2$ is shown in Fig. 16. Here positive and negative values of the argument were substituted in the equation of the parabola to find the corresponding values for y.

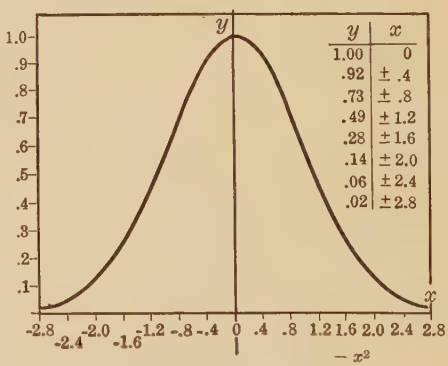


FIG. 17. Graph of $y = e^{\frac{-x^2}{2}}$

The *normal probability curve*, which will be used a great deal in the subsequent work, may be written in the form $y = e^{\frac{-x^2}{2}}$,

where e is the base of the Napierian system of logarithms and is equal to 2.71828. The curve in Fig. 17 has been plotted from the series of values furnished at the right. These values could be calculated directly, but are readily obtained from tables already prepared. It is evident that the same positive and negative values of the argument give only one value for the function, so that the curve is symmetrical about the y-axis. It is also to be noted that the vertical scale unit was not taken equal to the horizontal one. The choice of scale units will of course in no way alter the properties of the curve and is largely a matter of taste unless the curve is to be "fitted" to a series of observations. (See Chapter XII, section 5.)

EXERCISES

1. Calculate the valuation per inhabitant from the following data, computing the per capita valuations to the nearest dollar. Make a table ruled up according to the specifications in section 2. The columns in the table will be (1) city, (2) population, (3) total valuation, (4) valuation per inhabitant, (5) rank.

City	Population in 1914 (Thousands)	Estimated Valuation of All Property Assessed (Thousands)
Baltimore	580	$723,800
Boston	734	1,489,609
Buffalo	454	494,200
Cleveland	639	756,831
Detroit	538	598,634
Indianapolis	259	363,414
Jersey City	294	257,645
Kansas City	282	371,191
Los Angeles	439	836,604
Milwaukee	417	511,721
Minneapolis	343	639,259
Newark	389	383,864
New Orleans	361	314,086
Pittsburgh	565	789,035
San Francisco	449	1,247,391
Seattle	313	473,175
St. Louis	735	1,125,309
Washington	353	538,390

2. Make a column diagram for the following data on centimeter or similar graph paper:

a. Each column represents a grade and is proportional in height to the membership of the grade.

b. Darken the upper part of each column to show the proportion of overage children in each grade.

c. Make the columns one centimeter wide and leave a one-half-centimeter space between columns.

d. Print the total membership over each column or use a scale at the left.

e. Put a suitable title at the bottom of the diagram.

NUMBER OF NORMAL AND OVERAGE PUPILS IN AN IDEAL SCHOOL IN WHICH AN 80 PER CENT PROMOTION RATE IS IN EFFECT

GRADE	TOTAL	NORMAL	OVERAGE
I	125	120	5
II	125	112	13
III	125	103	22
IV	125	92	33
V	124	82	42
VI	121	73	48
VII	109	63	46
VIII	85	55	30

3. Plot the following pairs of scores for quality and speed on the Ayres Handwriting Scale.

Q. 42, 31, 65, 59, 38, 62, 35, 47, 57, 67, 51, 42, 34, 29, 63
S. 94, 91, 87, 81, 80, 78, 75, 74, 75, 73, 70, 68, 61, 43, 75

4. Make histograms for the distributions found in Exercise 3 of Chapter II.

5. Make histograms for the two spelling distributions of Exercises 6 and 7 of Chapter II.

6. Construct graphs for the cumulative frequency distributions given by Exercises 4 and 8 of Chapter II.

7. Plot the straight lines,

(*a*) $y = 3x - 7$, (*b*) $y = 2x + 6$, (*c*) $x = 3y - 4$.

8. Plot the curves,

(*a*) $y = 3x^2 + 2x - 1$, (*b*) $y = 4x^3 - 6x^2 + 2x + 3$, (*c*) $y = 10e^{\frac{-x^2}{2}}$.

(Make use of the values given in section 7.)

CHAPTER IV

LOGARITHMS

1. Introductory

For most computations it is best to use a calculating machine, but for students such aids are frequently out of the question. They must often resort to ordinary arithmetic, slide rules, or logarithms in working out statistical problems. In dealing with classroom exercises and even extended problems such as those arising in connection with a thesis, logarithms will be found to be extremely convenient and accurate. The present chapter is therefore devoted to a brief account of their nature and use.

The student who is familiar with logarithms may omit this chapter, but it frequently happens that one needs to review this subject. The present material may then serve not only as a short introduction for the student who knows nothing of logarithms, but also as a convenient reminder of some of the things once known but forgotten.

2. Arithmetical and Geometrical Progressions

An arithmetical progression is a succession of terms such that each term differs from that immediately preceding it by a constant known as the common difference. Show that the following are examples of such arithmetical progressions or series:

	Arithmetical Progression	Difference, d
a.	$1, 2, 3, 4, 5, 6, \cdots$	$+1$
b.	$16, 14, 12, 10, 8, 6, \cdots$	-2
c.	$6, 11, 16, 21, 26, 31, 36, \cdots$	$+5$
d.	$2\frac{1}{2}, 3\frac{3}{4}, 5, 6\frac{1}{4}, 7\frac{1}{2}, \cdots$	$+1\frac{1}{4}$
e.	$-5, -3, -1, +1, +3, +5, \cdots$	$+2$
f.	$a, a+d, a+2d, a+3d, \cdots$	$+d$

A geometrical progression is a series of terms such that each term is the product of the preceding term by a constant known as the ratio. Examples of such progressions are as follows:

	Geometrical Progression	Ratio, r
a.	$1, 2, 4, 8, 16, 32, \cdots$	2
b.	$100, -50, 25, -12.5, 6.25, \cdots$	$-.5$
c.	$5, \frac{5}{3}, \frac{5}{9}, \frac{5}{27}, \frac{5}{81}, \cdots$	$\frac{1}{3}$
d.	$a, ar, ar^2, ar^3, ar^4, \cdots$	r

The abbreviation for an arithmetical progression is *A.P.* and for a geometrical progression is *G.P.*

The *arithmetical mean* of a series is obtained by dividing the total of the numbers by the number of items in the series. Thus in the series 1, 2, 3, 4, 5, 6, 7, the mean is $28/7 = 4$. A general procedure for finding the mean of any arithmetical series may be shown as follows: Let the first term and the difference be any algebraic numbers denoted by a and d and let n be a positive integer representing the number of terms. We may then write

Number of term:	1	2	3	$\cdots$	n
Progression:	a	$a+d$	$a+2d$	$\cdots$	$a+(n-1)d$

The last term, or l, is clearly given by the formula

$$l = a + (n-1)d. \tag{1}$$

If s denotes the sum of the n terms in such a progression, this sum written in natural and in reverse order will give

$$s = a + [a+d] + [a+2d] + \cdots [a+(n-1)d]$$

and $$s = l + [l-d] + [l-2d] + \cdots [l-(n-1)d].$$

Adding these two equations, member by member, we find that

$$s = \frac{n(a+l)}{2}. \tag{2}$$

The arithmetical mean, *A.M.*, is therefore given by

$$A.M. = \frac{s}{n} = \frac{a+l}{2}. \tag{3}$$

Applying this formula to the *A.P.* 6, 11, 16, 21, 26, 31, we obtain $A.M. = \frac{6+31}{2} = 18.5$.

If three numbers form a *G.P.* the middle number is called the *geometrical mean* of the other two and is obtained by extracting the square root of their product. This follows at once from the general form of a *G.P.*: $a, ar, ar^2 \cdots ar^{n-1} = l$. For any two numbers a and b, therefore, the geometrical mean is given by the formula

$$G.M. = \sqrt{ab}. \tag{4}$$

EXAMPLE. The *G.M.* of 1 and 9 is $\sqrt{1 \times 9} = 3$, that is, 1, 3, 9 are in a *G.P.* with ratio 3.

Insert four geometric terms between 18 and $\frac{2}{27}$. The first term $a = 18$, $l = \frac{2}{27}$, and $n = 4 + 2 = 6$. Since $l = ar^{n-1}$ we have $\frac{l}{a} = r^{n-1} = \frac{1}{3^5}$, whence $r = \frac{1}{3}$. The required terms are therefore 6, 2, $\frac{2}{3}$, and $\frac{2}{9}$.

3. The Invention of Logarithms

The most important discovery in the development of mathematical computation was the invention of logarithms by John Napier, Baron of Merchiston of Scotland (1550–1617). The principle underlying his invention may be explained in terms of arithmetical and geometrical progressions.

Let such a pair of associated series be given as follows:

A.P.	0	1	2	3	4	5	6	7	8	9	10
G.P.	1	2	4	8	16	32	64	128	256	512	1024

The product of any two numbers in the second line of numbers (*G.P.*) may be found by adding the corresponding numbers in the *A.P.*, finding this sum in the *A.P.*, and finally taking the corresponding number in the *G.P.* line as the required answer. Thus the product 4×128 may be found by adding 2 and 7 (the numbers in the *A.P.* corresponding), finding their sum (9) in

the *A.P.*, and then the corresponding number (512) in the *G.P.*, this being the required product. The time-saving principle illustrated by this method is that the process of multiplication is replaced by that of addition.

It is apparent that series such as the above furnish only a few of the possible products which might be required. The system needs, therefore, to be extended. In addition to continuing the progressions at either end, Napier inserted terms as illustrated by the following series:

A.P.	0	.5	1	1.5	2	2.5	3	3.5
G.P.	1	$\sqrt{2}$	2	$\sqrt{8}$	4	$\sqrt{32}$	8	$\sqrt{128}$
	1	1.41	2	2.83	4	5.66	8	11.31

This amounts to inserting arithmetical and geometrical means between the original terms.

The above series are *tabular representations* of the function $y = 2^x$, where x denotes the numbers in the *A.P.*, and y the numbers in the *G.P.* The number 2 is called the base and x is said to be the logarithm of y to the base 2. *The logarithm of a number is thus the exponent to which a fixed number, called the base, must be raised to equal the given number,* or, if $y = b^x$, then x is the logarithm of y to the base b, or $x = \log_b y$.

If 2 is the base, $\log_2 64 = 6$, because $2^6 = 64$; if 8 is the base, $\log_8 64 = 2$, because $8^2 = 64$. The number of possible bases is clearly infinitely large.

The invention of logarithms by Napier stimulated an Englishman by the name of Henry Briggs to work out a system of logarithms to the base 10. Between the years 1617 and 1628, Briggs and others completed tables of logarithms up to 100,000 carried out to fourteen decimal places. Many other tables have since been computed, one of the most complete being a 20-place table carried out by Mr. A. J. Thompson* under the direction of Professor Pearson.

* A. J. Thompson, Logarithmetica Britannica, being a Standard Table of Logarithms to Twenty Decimal Places. Cambridge University Press, London, 1924.

The present chapter will deal entirely with the Briggs logarithms with a base of 10. Before considering their use, however, a brief review of the laws of exponents will be given.

4. Laws of Exponents

The symbol a^n is used to represent the product of a to n equal factors, or $a^n = a \cdot a \cdot a \cdot a \cdots$ to n factors where n is a positive integral (whole) exponent. Certain fundamental laws for such exponents may now be given as follows:

I. $(a^m)(a^n) = a^{m+n}$; for example, $(10^2)(10^3) = 10^{2+3} = 10^5$.

This follows at once from the fact that

$$(a^m)(a^n) = (a \cdot a \cdot a \cdots \text{to } m \text{ factors})(a \cdot a \cdot a \cdots \text{to } n \text{ factors})$$
$$= a \cdot a \cdot a \cdot a \cdots \text{to } (m+n) \text{ factors.}$$

The remaining laws are proved in a similar way.

II. $\dfrac{a^m}{a^n} = a^{m-n}$; for example, $\dfrac{8^6}{8^4} = 8^{6-4} = 8^2$.

III. $(a^m)^n = a^{mn}$; for example, $(20^2)^3 = 20^6$.

IV. $(ab)^n = a^n b^n$; for example, $(3 \times 4)^2 = 3^2 \times 4^2$.

$$\left(\frac{2}{3}\right)^3 = \frac{2^3}{3^3}.$$

The above laws also hold when the exponents are any positive or negative integral or fractional numbers. Fractional and negative exponents are defined as follows:

$$a^{\frac{m}{n}} = \sqrt[n]{a^m}; \text{ for example, } 8^{\frac{2}{3}} = \sqrt[3]{8^2} = 4.$$

$$a^{-n} = \frac{1}{a^n}; \text{ for example, } 16^{-2} = \frac{1}{16^2} = \frac{1}{256}.$$

$$16^{-\frac{1}{2}} = \frac{1}{\sqrt{16}} = \frac{1}{4}.$$

If $a^{-n} = \dfrac{1}{a^n}$, it follows that $a^{n-n} = a^0 = 1$.

Thus any number to the zero power is equal to one. This is an important law and should be remembered.

Some further illustrations of the above laws are as follows:

$$\frac{16^{\frac{1}{4}}}{16^{\frac{1}{2}}} = 16^{\frac{1}{4}-\frac{1}{2}} = 16^{-\frac{1}{4}} = \frac{1}{\sqrt[4]{16}} = \frac{1}{2}.$$

$$(27^2)^{\frac{1}{3}} = \sqrt[3]{(27)^2} = 9.$$

$$(4 \times 9)^{-\frac{1}{2}} = 36^{-\frac{1}{2}} = \frac{1}{\sqrt{36}} = \frac{1}{6}.$$

$$\left(\frac{2}{3}\right)^{-\frac{1}{2}} = \left(\frac{3}{2}\right)^{\frac{1}{2}} = \sqrt{1.5} = 1.225.$$

$$\frac{5^0}{5^2} = \frac{1}{5^2} = \frac{1}{25}.$$

5. Laws of Logarithms

From the definition of a logarithm and the laws of exponents, the basic principles for logarithmic computation may be expressed as follows:

I. *The logarithm of a product is equal to the sum of the logarithms of the factors, or*

$$\log_b MN = \log_b M + \log_b N.$$

Proof. Let $x = \log_b M$ and $y = \log_b N$. Then $b^x = M$ and $b^y = N$ (from the definition in section 3) and $MN = b^{x+y}$, or $\log_b MN = x + y = \log_b M + \log_b N$. The proofs for the remaining laws are similar.

II. *The logarithm of a quotient is equal to the logarithm of the dividend minus the logarithm of the divisor, or*

$$\log_b \frac{M}{N} = \log_b M - \log_b N.$$

III. *The logarithm of the nth power of a number is n times the logarithm of the number, or*

$$\log_b M^n = n \log_b M.$$

IV. *The logarithm of the nth root of a number is one-nth of the logarithm of the number, or*

$$\log_b \sqrt[n]{M} = \frac{1}{n} \log_b M.$$

6. The Briggs System of Logarithms

Returning to the Briggs system of logarithms we may note that the logarithm of any number N to the base 10 is the exponent x to which 10 must be raised to produce the number N. Thus, if

$$x = \log_{10} N,$$

then

$$10^x = N.$$

Inasmuch as 10 is always the base here considered we may hereafter write more briefly,

$$x = \log N.$$

From the above definition we may write down the logarithms of certain numbers at once as shown in Table 9.

Table 9. Showing the Logarithms of Numbers which are Multiples of 10

Number	Logarithm	Authority
100,000.	5	$10^5 = 100{,}000.$
10,000.	4	$10^4 = 10{,}000.$
1,000.	3	$10^3 = 1{,}000.$
100.	2	$10^2 = 100.$
10.	1	$10^1 = 10.$
1.	0	$10^0 = 1.$
.1	−1	$10^{-1} = .1$
.01	−2	$10^{-2} = .01$
.001	−3	$10^{-3} = .001$
.0001	−4	$10^{-4} = .0001$
.00001	−5	$10^{-5} = .00001$

The logarithm of a number between 100 and 1000 will evidently be somewhere between 2 and 3, that is, some fractional exponent. The number having the logarithm 2.5, for example, may be found by taking the geometric mean of 100 and 1000, or $\sqrt{100{,}000} = 316.2$. We may then write $\log 316.2 = 2.5$.

It is evident that logarithms consist of an integral and a decimal part, the former being called the *characteristic* and the latter the *mantissa* of the logarithm. Thus for the logarithm of 316.2 the characteristic is 2 and the mantissa is .5.

Very complete tables of mantissas have been computed as described above and conveniently tabled for use. The characteristic, it will be noted, may always be obtained by inspection.

In order to illustrate the procedure in finding the complete logarithm a four-place table of mantissas is given on pages 60 and 61. Let the logarithm of 43.2 be required. Since this number lies between 10 and 100 its logarithm will be between 1 and 2 and hence the characteristic is 1.

The mantissa, or decimal part, is found by looking down the column under N for the figures 43 and then proceeding to the right until the column headed 2 is reached. The number found is 6355. The decimal points have been omitted in the table, so that the complete logarithm is $1 + .6355$, or $\log 43.2 = 1.6355$.

If the number had been 4.32, the characteristic would have been zero and the mantissa the same as before. Therefore, $\log 4.32 = 0.6355$. This result is evident from the laws of exponents, for if

$$10^{1.6355} = 43.2,$$

then

$$10^{1.6355} \div 10 = 4.32,$$

or

$$10^{0.6355} = 4.32.$$

For the logarithm of .432, the characteristic will be -1 and the mantissa will again be equal to $+ .6355$. Instead of adding these two values directly, however, it is found more convenient to keep the mantissa positive and write

$$\log .432 = 9.6355 - 10,$$

by adding and subtracting 10 from the characteristic.

The general rule for determining the characteristic of a logarithm may now be stated as follows: *The characteristic of a number greater than 1 is one less than the number of digits to the left of the decimal point; while the characteristic for a number less than 1 is negative and one greater (numerically) than the number of zeros between the decimal point and the first significant figure.*

In looking up the mantissa of a number the rule is to neglect the decimal point and find the nearest mantissa for the given

sequence of digits. A more accurate method will be shown in section 7, where linear interpolation is presented.

The logarithms of the following numbers should now be verified by these rules and Table 10.

$\log 6.37 = 0.8041$	$\log .00004 = 5.6021 - 10$
$\log .0637 = 8.8041 - 10$	$\log 1910 = 3.2810$
$\log .00637 = 7.8041 - 10$	$\log 20000 = 4.3010$
$\log 1.01 = 0.0043$	$\log 2 = 0.3010$
$\log .001 = 7.0000 - 10$	$\log .999 = 9.9996 - 10$

A few short calculations may now be illustrated by the use of logarithms. Let the product 6.37×1910 be required. By the first law of the preceding section,

$$\log (6.37 \times 1910) = \log 6.37 + \log 1910$$
$$= 0.8041 + 3.2810 = 4.0851.$$

The number corresponding to the logarithm 4.0851 is clearly between 10,000 and 100,000, and the sequence of the digits is determined by the mantissa .0851. The nearest mantissa in Table 10 is .0864, corresponding to the number 122, so that the required product to three figures is 12,200. By direct multiplication the product is 12,166.70.

The steps in the above calculation were as follows:

1. Finding the logarithms of the factors (.8041 and 3.2810),
2. Adding these logarithms (4.0851),
3. Looking for the number (N) corresponding to the mantissa of the sum of the logarithms (122 corresponds to .0864), and
4. Determining the number of places in the result by noting the characteristic of the sum of the logarithms (characteristic 4 gives five digits before decimal point), and supplying zeros for the missing digits. (Answer is 12,200.)

Next let the quotient $\frac{.0437}{6920}$ be required. By the second law of logarithms, log quotient = log .0437 − log 6920, or $(8.6405 - 10) - 3.8401 = 4.8004 - 10$. The reason for adding and subtracting 10 for negative characteristics now becomes apparent, for the

subtraction may be made continuous; that is, on reaching the decimal point, 1 may be borrowed from the 8, which is positive. The difference is therefore $4.8004 - 10$. Looking in the table for the mantissa nearest .8004 we find .8007, which corresponds to the sequence of digits 632. The characteristic $4 - 10$, or -6, shows that five zeros must follow between the decimal point and the first significant figure in the number. The required quotient is therefore .00000632. By arithmetical calculation we obtain $.000006315+$.

The great convenience of logarithms is shown especially in raising a number to a given power. If $(.642)^6$ be required, the third law of logarithms may be applied, and we find that

$$\begin{aligned} \log (.642)^6 = 6 \log .642 &= 6(9.8075 - 10) \\ &= 58.8450 - 60 \\ &= 8.8450 - 10. \end{aligned}$$

The nearest mantissa is .8451 for $N = 700$, and the characteristic is -2. The answer is therefore .0700. By multiplying out $(.642)(.642) \cdots$ to six factors we obtain .07002.

By applying the fourth law, $\sqrt[10]{.777}$ may be found as follows:

$$\begin{aligned} \log \sqrt[10]{.777} = \tfrac{1}{10} \log .777 &= \tfrac{1}{10}(9.8904 - 10) \\ &= \tfrac{1}{10}(99.8904 - 100) \\ &= 9.98904 - 10. \end{aligned}$$

The required root is therefore .975.

7. Interpolation

A graph of the logarithm function $y = \log_{10} N$ may be made by plotting a few of the values from Table 10. (See Fig. 18.)

The logarithm of 7 is given by the ordinate .8451, while the logarithm of 8 is represented by $y = .9031$. If the logarithms between 7 and 8 were unknown, an approximation to the logarithm of 7.5 could be obtained by assuming that the function is a straight line over this interval and taking the ordinate at 7.5

as the required logarithm. Graphically, this amounts to measuring the ordinate PQ shown in Fig. 18. Arithmetically, the procedure is to take half the sum of the logarithms of 7 and 8, or .8741. Reference to the table, however, gives log 7.5 = .8751, so that there is an error of .001 in this case.

The above method is known as *linear interpolation* and is extremely useful in case the interval over which the interpolation is carried is *small*. In such cases the function will be so nearly linear that only a slight error will result. Values of the function between those given in the table may be found, and hence a greater degree of accuracy may be obtained than in the tabled entries.

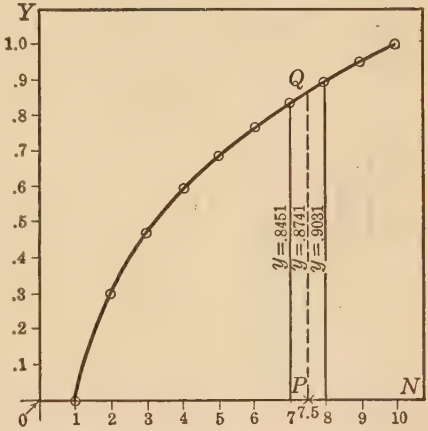


FIG. 18. Graph of $y = \log n$ illustrating linear interpolation for log 7.5

Thus, if the logarithm of 7.637 be required, the logarithms of 7.63 and 7.64 may be found in Table 10 and the extra amount for .007 found by interpolation. An enlarged portion of the graph is shown in Fig. 19. The difference between the logarithm for 7.64 and 7.63 is .0006 and is known as the *tabular difference*. From similar triangles it is now apparent that $\frac{c}{.007} = \frac{.0006}{.010}$, or $c = .7(.0006) = .0004$, where c is the *correction* to be added to the lower tabular value. The logarithm of 7.637 is therefore .8825 + .0004 = .8829. (From a seven-place table the logarithm is .8829228.)

The labor of computing each correction is saved by using a table of *proportional parts* shown at the right of the main figures in Table 10. In finding the logarithm of 7.637, for example, it is only necessary to look up the logarithm of 7.63, move out along

this line to the value 7 at the top or bottom of the proportional parts, and read off the entry 4. This last result is to be added to the fourth place of .8825, giving .8829 as before. In a similar way, the logarithms of 6.349 and .04233 are 0.8027 and 8.6266 − 10, respectively.

The table of proportional parts is also useful in looking up the number corresponding to a given logarithm. This may be illustrated by the following problem:

Find the product of .7437 and 3.242.

$$\begin{aligned} \log .7437 &= 9.8714 - 10 \\ \log 3.242 &= 0.5108 \\ \log \text{prod.} &= 10.3822 - 10 \end{aligned}$$

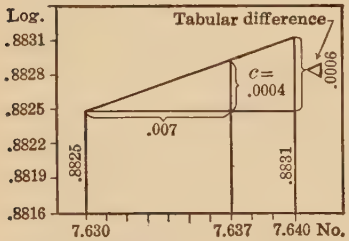


FIG. 19. Showing linear interpolation for log 7.637

The nearest mantissa smaller than .3822 is .3820, which corresponds to the number 241. The difference .0002 is now found in the proportional parts on the same line and by moving up to the top is found to correspond to 1. This last result should be adjoined to the three figures already found, giving as the required number 2.411. The whole procedure will become clearer if the logarithm of 2.411 is now worked out as shown in the paragraph above.

The method of linear interpolation will be sufficiently accurate for small differences in logarithms and similar functions where one (and possibly two) places beyond those given in the table are required. Thus with Table 10 linear interpolation is adequate for the logarithms of four-place numbers, and with a five-place table such as Taylor's* similar interpolation gives logarithms of five-place numbers.

More exact methods of interpolation are often required in advanced statistical work, but the formulas become quite com-

* Taylor, Five-Place Logarithmic and Trigonometric Tables. Ginn and Company. This table is especially recommended on account of its excellent physical make-up and the thumb index with which it is provided.

plicated and are used so seldom in elementary work that they are omitted here. For a clear account the student should consult Forsyth,* and for more advanced treatment an excellent work by Whittaker and Robinson.†

8. Some Additional Problems

It should be noted that the operations of addition and subtraction of numbers cannot be carried out by logarithms. Thus, if the problem to be worked out is

$$\frac{(6.743)(89.24) - (36.5)}{475},$$

this must be broken up into two parts which are worked separately by logarithms and combined only when the final answers are obtained. The work will then be as follows:

$$\begin{array}{rl} \log 6.743 = & 0.8289 \\ \log 89.24 = & 1.9506 \\ \hline \log \text{prod.} = & 2.7795 \\ \log 475 = & 2.6767 \\ \hline \log \text{quot.} = & 0.1028 \\ \therefore \text{quot.} = & 1.267 \end{array} \qquad \begin{array}{rl} \log 36.5 = & 11.5623 - 10 \\ \log 475 = & 2.6767 \\ \hline \log \text{quot.} = & 8.8856 - 10 \\ \therefore \text{quot.} = & .07684 \end{array}$$

The required answer is therefore $1.267 - .077 = 1.190$. As we shall see in the next chapter, such a result should not be carried beyond four figures.

In subtracting the logarithm of 475 from log 36.5 it will be noted that 10 has been added to and subtracted from the characteristic of the latter in order to facilitate the final subtraction of the logarithms.

A typical problem that occurs in statistical calculation is of the form

$$S.D. = \left[\sqrt{\frac{S}{N} - C^2}\right]h, \quad \text{for example,} \quad \left[\sqrt{\frac{3483}{794} - \left(\frac{37}{794}\right)^2}\right]5.$$

* Forsyth, Mathematical Analysis of Statistics, chap. iii. Wiley, 1924.

† Whittaker and Robinson, The Calculus of Observations, D. Van Nostrand Company, 1924.

TABLE 10. FOUR-PLACE LOGARITHMS OF NUMBERS

N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7
N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

The proportional parts are stated in full for every tenth at the right-hand side. The logarithm of any number of four significant figures can be read directly by adding the proportional part corresponding to the fourth figure to the tabular number corresponding to the first three figures.

TABLE 10. FOUR-PLACE LOGARITHMS OF NUMBERS (CONTINUED)

N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4
N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

With a machine or even by long-hand arithmetic, this result may be worked out quite rapidly if a good table of squares is used. The logarithmic calculation is a little awkward on account of the subtraction under the radical, but inasmuch as some students find it convenient the method may be illustrated as follows:

$$\begin{aligned} \log 3483 &= 3.5420 \\ \log 794 &= 2.8998 \\ \hline \log \frac{S}{N} &= 0.6422 \\ \therefore \frac{S}{N} &= 4.387 \end{aligned} \qquad \begin{aligned} \log 37 &= 11.5682 - 10 \\ \log 794 &= 2.8998 \\ \hline \log C &= 8.6684 - 10 \\ \log C^2 &= 17.3368 - 20 \\ \therefore C^2 &= .002 \end{aligned}$$

$$\frac{S}{N} - C^2 = 4.385$$

$$\log\left(\frac{S}{N} - C^2\right) = 0.6420$$

$$\log\sqrt{\frac{S}{N} - C^2} = 0.3210$$

$$\begin{aligned} \log h &= .6990 \\ \hline \log S.D. &= 1.0200 \end{aligned} \qquad \therefore S.D. = 10.47$$

Another calculation may be illustrated by the formula

$$r = \frac{S}{N\sigma_x\sigma_y}. \quad \text{For example, } r = \frac{743.2}{682(2.673)(2.794)}.$$

This is readily adapted to logarithmic work:

$$\begin{aligned} \log 682 &= 2.8338 \\ \log 2.673 &= 0.4270 \\ \log 2.794 &= 0.4462 \\ \hline \log \text{prod.} &= 3.7070 \end{aligned} \qquad \begin{aligned} \log 743.2 &= 12.8711 - 10 \\ \log \text{prod.} &= 3.7070 \\ \hline \log r &= 9.1641 - 10 \\ \therefore r &= .1459 \end{aligned}$$

A final problem may be worked out in the case of the geometric mean of several quantities where

$$G.M. = \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdots X_n}$$

for example, $G.M. = \sqrt[5]{27.4 \times 29.5 \times 28.3 \times 29.2 \times 29.9}$

We therefore have:

$$\begin{aligned}
\log 27.4 &= 1.4378 \\
\log 29.5 &= 1.4698 \\
\log 28.3 &= 1.4518 \\
\log 29.2 &= 1.4654 \\
\log 29.9 &= 1.4757 \\
\hline
\log \text{prod.} &= 7.3005 \\
\tfrac{1}{5} \log \text{prod.} &= 1.4601 \\
\therefore\ G.M. &= 28.8
\end{aligned}$$

EXERCISES

1. Find the logarithms of the following numbers by a four-place table. Check your results by referring to a five-place table: 634.2, 59.61, 1.722, .004359, .1166, .00004795, 5566., 6234000.

2. Work out the following operations by logarithms:

$$(1)\ \frac{.6432 \times .03475}{6.742}, \quad (2)\ \sqrt[6]{\frac{437.1}{3622.}}, \quad (3)\ \frac{1}{\sqrt{472 \times 347}},$$

$$(4)\left[\frac{(.3472)^2(.6745)^4}{(1.342)^3}\right]^2, \quad (5)\ \sqrt[4]{67 \times 68 \times 69 \times 70}.$$

Ans. (1) .003315, (2) .7030, (3) .00247, (4) .0001066, (5) $\begin{cases} 68.49. \\ 68.4875. \end{cases}$

3. Calculate the standard deviations with the following data, using the formula $S.D. = \left[\sqrt{\frac{S}{N} - C^2}\right] h$.

	S	N	C	h	$S.D.$ (*Ans.*)
(1)	4732	462	.0123	5	16.0
(2)	1692	192	1.1340	3	8.23
(3)	1573	641	.843	0.25	0.330

4. Compute the correlations for the data below, using the formula $r = \frac{a}{\sqrt{bc}}$.

	a	b	c	r (*Ans.*)
(1)	176	235	182	.851
(2)	234	234	259	.951
(3)	193	291	279	.677
(4)	− 64.2	173.3	1892	− .112
(5)	831	831	831	1.000

5. Compute the geometric means for the following series:
a. 169, 171, 165, 168, 173, 175, 170. (170.1. *Ans.*)
b. 33.1, 34.2, 33.4, 34.5, 33.6, 34.7, 34.8, 33.9. (34.0. *Ans.*)

6. Calculate $r_{12.3}$ by the following formula:

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}}$$

Use Holzinger's* Table VII for $\log \sqrt{1 - r^2}$.

	r_{12}	r_{13}	r_{23}	$r_{12.3}$ (*Ans.*)		r_{12}	r_{13}	r_{23}	$r_{12.3}$ (*Ans.*)
(1)	.82	.16	.17	.815	(4)	.431	.327	.214	.391
(2)	.09	.16	.17	.065	(5)	.647	.832	.725	.115
(3)	.80	.80	.80	.444	(6)	.932	.327	.214	.934

7. Using the formula

$$1 - R_{1(234)}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2),$$

work out the values of $R_{1(234)}$ with the aid of Holzinger's Table VI.

	r_{12}	$r_{13.2}$	$r_{14.23}$	$R_{1(234)}$ (*Ans.*)
(1)	.791	.620	.474	.906
(2)	.833	.695	.347	.928
(3)	.755	.062	− .007	.756
(4)	.815	.742	.676	.958

* Karl J. Holzinger, Statistical Tables for Students in Education and Psychology. The University of Chicago Press, 1925.

CHAPTER V

ERRORS IN CALCULATION AND MEASUREMENT

1. Accuracy in Statistical Method

In dealing with statistical material it is desirable to recognize very early the importance of accuracy not only in the calculations which need to be performed but in the data themselves. The student should train himself to be accurate in his computations and to employ adequate checks wherever possible. He should also be cautious as to accuracy of the data which he is using, in order to safeguard against making unwarranted conclusions from the results obtained.

Actual blunders in calculation can best be obviated by extreme care and adequate methods of checking all of the computations. Even with such mistakes eliminated, however, it is necessary to be cautious regarding the number of places to use in order to obtain a result to a given degree of accuracy. The distinction between different types of error is also important. For these reasons the present chapter will be devoted to some of the simplest principles involved in errors of calculation and measurement.

2. Absolute and Relative Errors

An *error* may be defined as the discrepancy between the obtained and the true values from a numerical process or measurement. If X_1 be an obtained value and X the true value, the difference, $E_1 = X_1 - X$, is known as the *absolute error*. The ratio of the absolute error to the true value, or E_1/X, is called the *relative error*. For example, suppose the true value of X is 67.5 inches, and measurements $X_1 = 66.9$ inches and $X_2 = 69.7$ inches have been made. The two absolute errors will be

$E_1 = -0.6$ and $E_2 = +2.2$, while the corresponding relative errors will be $-.01$ and $+.03$, or -1 per cent and $+3$ per cent.

Whenever values are obtained from the measurements of some continuous variables such as height, they can never be exact nor can their true value ever be determined. All such values, including the errors themselves, must be approximations. The best that can be done is to measure to a certain degree of accuracy, take the average of a number of observations as an approximation to the true value, and consider the variations from this result as errors. Thus suppose a stick is measured ten times to the nearest millimeter and the following observations are recorded: 57, 58, 58, 56, 57, 60, 57, 55, 56, 56. Their average, or 57, might be taken as the true or most typical value, and the variations 0, $+1$, $+1$, -1, 0, $+3$, 0, -2, -1, -1 would be considered as absolute errors although they are themselves only approximations to the true errors.

In case we are dealing with a discrete series such as the numbers of pupils in various school grades the resulting observations of grade size may be considered as exact. It should be noted, however, that the unit of tabulation in such a series is the pupil, and that these units are equal to one another only in a very limited sense, that is, as human entities.

3. Biased and Unbiased Errors

Errors which tend to compensate or offset one another in the long run are known as *unbiased* or *compensating errors*. A good example is furnished by the *rounding off* of numbers to a smaller number of places as in Table 11.

In rounding off the numbers to the nearest thousand, figures less than 500 are discarded and those greater than 500 are considered as 1000. If figures had occurred at exactly 500, they would have been equally divided above and below, or in case of a single such number, 1000 would have been added. In the table on page 67 the "errors" in rounding were -347, -143, $+365$, $+228$,

TABLE 11. AVERAGE DAILY SCHOOL ATTENDANCE IN THE UNITED STATES, 1870–1910

YEAR	CHILDREN IN A.D.A.	THOUSANDS OF CHILDREN IN A.D.A.
1870	4,077,347	4,077
1880	6,144,143	6,144
1890	8,153,635	8,154
1900	10,632,772	10,633
1910	12,827,307	12,827
Total	41,835,204	41,835
Average	8,367,041	8,367

and -307, with an algebraic total of -204. Even with so short a series the total error was relatively small, and for a longer list of numbers it would tend to become less because of the random distribution of digits greater and less than five.

Unbiased errors are very important in the theory of averages because their balancing effect will tend to make the average more accurate than the original numbers. Thus, in Table 11, the absolute error in the rounded average is only 41, which is much less than that of any individual item. As a caution it should be added that the above short series happened to illustrate the above principles and was therefore chosen. In general a longer series would be required to secure any considerable balancing of errors.

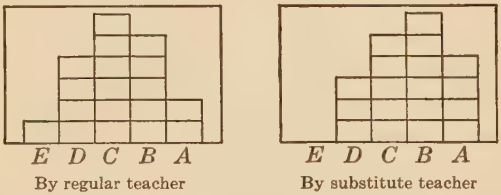


FIG. 20. Distribution of grades of a class of pupils by a regular teacher and by a substitute

Based on one month's class work and five tests

Biased errors are those which do not tend to neutralize one another but accumulate in such a way as to produce a relatively large error in the total or average. Such errors are illustrated in Fig. 20 by the marks of two teachers.

If biased errors are present in a series of observations, the average will tend to be as inaccurate as the individual measurements upon which it is based. Suppose that a meter stick is one centimeter shorter than the standard. All measurements with it will have a relative error of 1 per cent in the same direction and the average will be likewise affected as illustrated in the following table:

TABLE 12. HYPOTHETICAL MEASUREMENTS WITH CONSTANT ERROR OF 1 PER CENT

OBSERVED MEASUREMENT	CONSTANT ERROR
69	+ .69
71	+ .71
69	+ .69
68	+ .68
73	+ .73
Total 350	+ 3.50
Average . . . 70	+ .70

4. SIGNIFICANT FIGURES

The digits in a numerical result which are known to be correct are called significant figures. Thus, if a measurement such as 39.6 mm. be made, it is assumed to be correct to the nearest tenth of a millimeter and is said to have three significant figures, the true value lying anywhere between 39.55 mm. and 39.65 mm. If the same result is expressed as .0396 meter, it is still to be considered as correct to three figures, the zero after the decimal point merely serving to fill a space. When zeros occur on the right of a series of digits the significant figures may be shown by the use of a decimal point. For example, a measurement such as 2600. is correct to four figures or is between 2599.5 and 2600.5, while 2600 is correct to only two figures and lies between 2550 and 2650. By way of further illustration, the following numbers would all be considered as correct to five significant figures: 47.234, .00036924, .0042000, 4349.0, 1000.0, 956340, 1.0000.

5. Arithmetical Computation with Rounded Numbers

Consider the following series of products with successively rounded values of $\pi = 3.1415927$ and $e = 2.7182818$, whose product, correct to eight significant figures, is 8.5397342.

$\pi \times e$	Product	Correct Value
(3.1415927)(2.7182818)	= **8.53973425942286**	8.5397342
(3.141593)(2.718282)	= **8.539735703226**	8.539734
(3.14159)(2.71828)	= **8.5397212652**	8.53973
(3.1416)(2.7183)	= **8.53981128**	8.5397
(3.142)(2.718)	= **8.539956**	8.540
(3.14)(2.72)	= **8.5408**	8.54
(3.1)(2.7)	= **8.37**	8.5
(3)(3)	= **9.**	9.

The bold-faced figures are those which agree with the correct values on the right when the remaining digits are consolidated. Thus the first product is correct to seven significant figures only, for if rounded one place further to the right there would have been an error of 1 in the seventh decimal place, that is, 8.5397343 instead of 8.5397342. Of the remaining products only three are correct to as many significant figures as occur in each factor, while three others are correct to one less figure. The table illustrates the rule that it is not safe to carry out the product of two such factors beyond the number of significant figures included in each.

The same principle may be illustrated in another way. Suppose that the product of 36.9 by 8.74 is required, both factors being correct to three significant figures. The obtained product is 322.506. The maximum product is 36.95×8.745, or 323.12775, while the minimum product is 36.85×8.735, or 321.88475. In this problem it is therefore doubtful whether the correct answer is 322 or 323. To give the result to two significant figures, as 320, would not be desirable, for both maximum and minimum products exceed the value. The answer 323 is to be preferred because it is nearer the average of the extreme products and therefore more probably correct than 322.

When several factors are involved the rounding errors will tend to offset one another, but in case they do not, the error in the product may be relatively great, as illustrated by the following example:

Rounded value	34	× .129	× 6.784	= 29.754624 = 30
Maximum value	34.5	× .1295	× 6.7845	= 30.311450 = 30
Minimum value	33.5	× .1285	× 6.7835	= 29.201272 = 29

The best answer for this particular problem is probably 30, but if a dozen more items were included the product might be given to one more significant figure than in the item with the least significant figures.

In the case of division, similar reasoning may be applied. Consider, for example, the quotient $8.47 \div 23 = .368$. The maximum and minimum values are $8.475 \div 22.5 = .377$ and $8.465 \div 23.5 = .360$, respectively. Here there is such wide variation in the third figure that it could not be safely retained in the answer. The quotient .37 is probably best as an average between the extreme values .38 and .36.

The general rule, then, is that *a product or a quotient with rounded numbers should not be written to more significant figures than occur in the item with the least significant figures.*

For addition and subtraction it may also be shown that the result should not be written to more decimal places than occur in the least accurate measurement. Thus the sum of 624.2 feet and 49.173 feet should be written $624.2 + 49.2 = 673.4$ feet, and the difference as $624.2 - 49.2 = 575.0$ feet. By the same reasoning as is applied in multiplication, it is evident that the maximum and minimum sums are 673.4235 and 673.3225, whereas the corresponding differences are 575.0775 and 574.9765.

When several items are added it is probably best to round them at once to the number of decimal places in the least accurate measurement. If compensating errors occur, they will offset one another in the rounding of the items just as well as if more places had been retained and the sum rounded. The problem on page 71 illustrates this principle.

HYPOTHETICAL PROBLEM ILLUSTRATING ROUNDING IN SUMS

ORIGINAL ITEMS	ROUNDED TO TWO DECIMAL PLACES	ROUNDED TO ONE DECIMAL PLACE
67.432	67.43	67.4
9.64	9.64	9.6
10.4	10.4	10.4
8.356	8.36	8.4
17.9	17.9	17.9
6.566	6.57	6.6
8.327	8.33	8.3
7.463	7.46	7.5
29.638	29.64	29.6
19.784	19.78	19.8
Total 185.506	185.51	185.5

It is readily verified that the maximum and minimum sums are 185.6145 and 185.3975. The answer 185.5 is therefore the best, and it may be obtained as well from the last column of figures as from the second where the items have been carried to two decimal places and the sum rounded to one. The second decimal place is very doubtful in any event, and it is hardly worth while to retain two places in the items in an attempt to secure a better approximation to the correct result than could be obtained by rounding off to one place.

6. LOGARITHMIC COMPUTATION WITH ROUNDED NUMBERS

Inasmuch as a good share of the students' calculations may be performed with the aid of logarithms it may be well to discuss briefly their use with rounded numbers. As an illustration let the product 3.47×8.96 be required. The maximum and minimum factors f_1 and f_2, their logarithms, and the resulting products may be set down as follows:

	f_1	f_2	LOG f_1	LOG f_2	LOG $f_1 f_2$	$f_1 f_2$
Maximum .	3.475	8.965	.5409548	.9525503	1.4935051	31.153
Actual . . .	3.47	8.96	.5403295	.9523080	1.4926375	31.091
Minimum .	3.465	8.955	.5397032	.9520656	1.4917688	31.029

The best or most probable answer is 31.1, which is the product of 3.47 and 8.96 carried to three significant figures, and in order to obtain it four-place logarithms are as satisfactory as the seven-place. The abbreviated computation would then be

$$\begin{aligned} \log 3.47 &= .5403 \\ \log 8.96 &= .9523 \\ \hline \log \text{prod.} &= 1.4926 \\ \therefore \text{prod.} &= 31.1. \end{aligned}$$

When four significant figures are involved, four-place logarithms may be employed, but a five-place table is much more convenient because no interpolation is necessary if the entries for N are given to four places. For example, the computation of the product 123.7 by 96.45 may be done in either of the following ways:

With a Four-Place Table and Interpolation	With a Five-Place Table (No Interpolation)
$\log 123.7 = 2.0923$	$\log 123.7 = 2.09237$
$\log 96.45 = 1.9843$	$\log 96.45 = 1.98430$
$\log \text{prod.} = 4.0766$	$\log \text{prod.} = 4.07667$
$\therefore \text{prod.} = 11{,}930.$	$\therefore \text{prod.} = 11{,}930.$

With a product such as 34.79 by 7643.29, the second factor should be consolidated to 7643 or 7643.3 and a five-place table of logarithms employed.

The general rule that will apply also in the case of division is that *when n is the least number of figures to which any of the items is correct, an n or at most an n + 1 place logarithm table should be used.*

In logarithmic calculation involving formulas the same general rule may be followed. Thus in the case of the functions $1 - r^2$ and $\sqrt{1 - r^2}$, which occur very frequently, three-place, four-place, or five-place logarithm tables will be ample when the values of r are given to two, three, and four places, respectively. The following calculation illustrates the variations which may occur in the numbers and logarithms:

r			$1-r^2$			Log $(1-r^2)$*		
Value	Min.	Error	Value	Max.	Error	Value	Max.	Error
.18	.175	.005	.9676	.9694	.0018	9.9857	9.9865	.0008
.50	.495	.005	.7500	.7550	.0050	9.8751	9.8779	.0028
.82	.815	.005	.3276	.3358	.0082	9.5153	9.5261	.0108
.99	.985	.005	.0199	.0298	.0099	8.2989	8.4739	.1750

It will be noted that when the value of r is given correct to two places, the logarithm of $(1-r^2)$ may have an error in the first, second, or third place, etc., depending upon the size of r. Three-place logarithms of $(1-r^2)$ would therefore be sufficient for such problems.

In a similar way it may be shown that while a product such as $[1-(.856)^2][1-(.943)^2]$ may have a rounding error of only .00035 there may be an error of .005 in its logarithm, due to an addition of the errors in the two factors, as shown below.

$$[1-(.8565)^2][1-(.9435)^2]=.02925,$$
$$[1-(.856)^2][1-(.943)^2]=.02960,$$
$$[1-(.8555)^2][1-(.9425)^2]=.02995.$$

Maximum rounding error = .00035.

$$\begin{array}{rr} \log[1-(.856)^2]=9.42694-10 & \log[1-(.8555)^2]=9.42833-10 \\ \log[1-(.943)^2]=9.04435-10 & \log[1-(.9425)^2]=9.04803-10 \\ \hline \log \text{ prod.}=8.47129-10 & \log \text{ prod.}=8.47636-10 \end{array}$$

Maximum error in log prod. = .00507.

The product to be chosen is surely right when written to one significant figure, as .03, but it might be correct to three significant figures, as .0296, on account of compensating errors.

In case it is desired to have the final answer for a problem correct to n significant figures, it is usually best to begin with the items correct to $n+1$ significant figures and use $n+2$ place tables in the computation.

*Karl J. Holzinger, Statistical Tables for Students in Education and Psychology. The University of Chicago Press, 1925. See Table VI for log $(1-r^2)$.

7. Errors in Educational Measurement

The errors discussed thus far have been due chiefly to the rounding of approximate measurements. They are not peculiar to any one field, but occur whenever measurements or observations are made and should be taken into account in the subsequent calculations. Being unbiased in character their effect upon the final result may be controlled by care in the arithmetical operations as described above. The present section will be concerned with errors which occur in the measurement of mental characters.

One difference between mental and physical measurements arises from the nature of the scales employed. Arithmetical ability, for example, is a very complex character and its resolution into component abilities such as those of addition and multiplication is at best a matter of convenience because each of these is a combination of still more specific abilities. A unit of such arithmetical ability can therefore never be quite the equivalent of another unit in the arithmetical scale in the same way that an inch of height is the equivalent of another inch of height. Even two problems alike in type and equally difficult for a large group may not be equally difficult for a single pupil. The inch, on the other hand, has the same significance for the individual measurement as in the group.

This lack of equivalence of test units is closely related to another difference between mental and physical scales. The complete measurement of a mental trait is probably impossible, because the test must always be based on a sampling of the total available material. Spelling ability, for example, may be measured by a number of well-known scales, but no single test nor the combination of several tests will give a complete measure of spelling ability. These tests, moreover, will be only roughly comparable because different words and methods of testing are employed. An approximate transmutation from one mental scale to another is always possible, but nothing approaching

the exactness with which inches may be converted into centimeters can probably ever be attained in the case of mental measurements.

The examiner or observer in giving a mental test may introduce certain errors by his failure to follow the uniform directions for the administration of the test. He may create an unfavorable mental attitude on the part of the pupils by hurrying them or urging them to be overcautious. In scoring the results he may make mistakes in using the key even with objective tests, or show poor judgment in rating the specimens in the case of product scales such as those for handwriting and composition.

Another source of error in mental measurement is associated with what Professor Pearson has called *static* as distinct from *dynamic* characters. The former include such physical traits as height and weight, the measurement of which is *direct* and does not depend upon the attitude of the person at the time of examination. Dynamic characters like lung capacity, strength of grip, or intelligence must be measured *indirectly* by some form of reaction, and therefore depend upon the bodily or mental fitness of the individual. The measurement of dynamic traits thus gives rise to a variability in reaction which may be called *response error* of the person tested.

It should be noted that when a pupil has been examined several times on equally difficult forms of a test, any change in his response may be due in part to the attitude of the examiner, to imperfections in the test material, to practice effect, to fluctuations in emotional status and fatigue, etc. Response error as measured by variation in score may thus be a combination of several of the types of error already discussed. Certain formulas which attempt to measure response variability freed from other error are presented in Chapter XIII, section 9.

As pointed out in the second section, the best approximation to the true value of any quantity is given by the average of a number of observations. For dynamic characters involving response error this conception of true value may be misleading.

If a person has been tested ten times on as many equivalent mental scales the average score may be the most typical one, but the highest score is likely to be the best representation of his true ability because on that performance there were fewer interfering factors which prevented him from doing himself full justice. The same argument might be made with regard to characters such as lung capacity. No matter how often the test is given the full lung capacity will never be registered, and the largest volume obtained may be considered as nearest the true result.

With standardized tests both the average (most typical) and the highest (nearest the true) scores will be useful, the former giving the best prediction as to future performance, and the latter the best indication of potential ability under most favorable conditions.

The above types of error in calculation and measurement may be briefly summarized as follows:

1. Unbiased or rounding errors to be taken into account in calculation.
2. Biased errors such as those found in teachers' marks.
3. Errors of the scale:
   a. Non-equivalent units or items;
   b. Inadequate sampling of available material.
4. Errors of the examiner:
   a. In giving the test;
   b. In appraising the results of the test.
5. Response error (or variation) of the examinee.

EXERCISES

1. Round off the following numbers to four significant figures: 35.675002, 846742., 390000., .6744898, .003674378.

2. If the numbers 39.2 and 18.3 are correct to three significant figures, justify the product 717. rather than 700.

HINT. Use maximum and minimum products.

3. Justify the quotient $18.3 \div 39.2 = 0.467$.

4. Show that the sum of 13.26818, 138.36, 78.423, 7238.4289, and 6.324 cannot be as large as 7474.82 or as small as 7474.79.

5. Show that the product of 34.68 and 4.6, carried to three digits, lies between 158 and 161.

6. Find the probable values of the following:

a. Sum of 27.843, 182.6, 5478.29, and 5.2777
b. Difference between 367.19 and 173.4395
c. Product of 897.5 and 0.08
d. Product of 37.846 and .0004
e. Quotient of 37.846 divided by .0004
f. Quotient of .0004 divided by 37.846

7. Calculate the following products, using Holzinger's Tables VI and VII and a five-place logarithm table of numbers. Repeat the calculations, rounding to four-place logarithms throughout, and compare results.

	Answers
a. $[1-(.346)^2][1-(.931)^2]$	$=.1173$
b. $[1-(.845)^2][1-(.674)^2]$	$=.1561$
c. $[1-(.113)^2][1-(.981)^2]$	$=.0372$
d. $\sqrt{[1-(.639)^2][1-(.846)^2]}$	$=.4101$
e. $\sqrt{[1-(.550)^2][1-(.947)^2]}$	$=.2683$
f. $\sqrt{[1-(.600)^2][1-(.400)^2]}$	$=.7332$

8. Discuss the theory of "most typical" and "nearest true" scores given in section 7. Do you agree with the distinction and use described by the author? If not, why not?

9. Can the measurement of mental abilities ever be made as exact as the measurement of physical objects? Explain.

10. Estimate the absolute and relative error made in measuring a person's height with an ordinary yardstick. Estimate the absolute and relative error made in measuring a person's intelligence, by a good group test and also by a good individual test. Use any data available to assist in these estimates.

CHAPTER VI

AVERAGES

1. Introductory

It has already been shown that the first step in making a long series of observations comprehensible is to arrange the data in the form of a frequency distribution. This enables one to see some of the more outstanding characteristics of the series at a glance, and at the same time makes subsequent calculations very much easier than they would have been with the data ungrouped.

The hypothetical distributions shown in Fig. 21 reveal certain important features by mere inspection. Curves (1) and (2) center about the value 15, which is a measure of type or *average*, but the first distribution is spread out more than the second. This second characteristic is known as *dispersion*, or *variability*. Distributions (3) and (4) are said to be *skewed*, the former negatively and the latter positively. Curve (5) is very steep (leptokurtic), whereas (6) is flat-topped (platykurtic). The first distribution, which is midway between the two, might be regarded as mesokurtic.

All these characteristics are very important in statistical analysis and they may all be quantitatively determined by appropriate formulas rather than by inspection of diagrams as illustrated in Fig. 21. In the present chapter methods will be presented for the calculation of several important averages, which include the mean, median, and mode. Measures of dispersion and skewness will be discussed in Chapter VII. The *kurtosis* of a distribution is so rarely studied that no formulas for its measurement are given in this text. Such formulas, however, may be found in Kelley's Statistical Method.

2. Calculation of the Mean

The most important and generally most reliable average happens also to be the best known. This is the *arithmetical mean.* It is defined simply as the sum of the values of the observations divided by their number, or by the formula

$$M = \frac{\Sigma X}{N}, \qquad \left\{\begin{matrix}\text{Mean for} \\ \text{ungrouped series}\end{matrix}\right\} \qquad (5)$$

where M is used to represent the arithmetical mean, X a value of the variable, and N the number of items. The symbol Σ

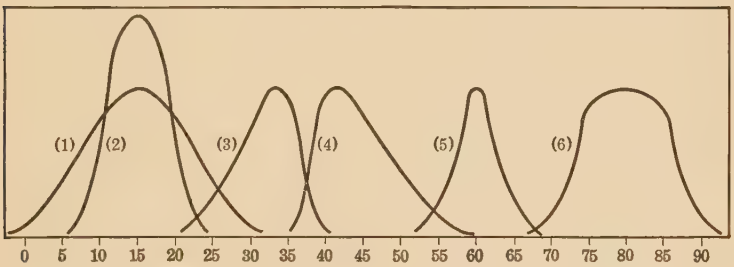


Fig. 21. Illustrating variations in central tendency, dispersion, skewness, and kurtosis

means "the sum of all quantities as follows," that is, the sum of all the X's. One property of the mean which follows at once from the above definition is that it is the magnitude each item would have if all items were the same size.

The calculation of the mean for ungrouped data is very simple. It is only necessary to add the items and divide by their number. For long series, however, this process becomes very tedious and errors in addition are likely to creep in. Calculation from the frequency distribution therefore becomes almost imperative with many items. The method will first be illustrated by the use of a short series which has been so selected that the attention of the student will first be directed to the method rather than to lengthy arithmetic. Needless to say, the series is too short for the average to be of any practical value.

Let the mean of the following scores be required: 97, 72, 63, 68, 93, 84, 79, 87, 56, 52, 64, 71, 75, 67, 64. The total of these items, ΣX, is 1092 and their mean is 72.8. This is the true mean within the limits of the accuracy of the data.

Next, assuming the scores are correct to the nearest unit only, we shall arrange them in a frequency distribution as follows:

Class	Frequency
89.5–99.5	2
79.5–89.5	2
69.5–79.5	4
59.5–69.5	5
49.5–59.5	2
	15

For purposes of calculation it is assumed that the frequencies are concentrated at the mid-points of the respective class intervals, such points being known as *class values* (Chapter II, section 8). The two top frequencies will thus contribute $2 \times 94.5 = 189.0$ to the total instead of $97 + 93 = 190$, and so on for the other classes, the complete calculation being

X	f	fX
94.5	2	189.0
84.5	2	169.0
74.5	4	298.0
64.5	5	322.5
54.5	2	109.0
	15	$1087.5 = \Sigma fX$

$$M = \frac{\Sigma fX}{N} = \frac{1087.5}{15} = 72.5.$$

It is evident that the sums ΣX and ΣfX differ by 4.5, a discrepancy which is due to the fact that the frequencies were taken at class values instead of at observed values. With a longer series and more class intervals the above discrepancy would be smaller, because the larger number of unbiased errors would tend to compensate, and with a narrower interval less variation from the class values would be possible. The means, it will be noted, differ by only 0.3 in spite of the short series and coarse classification. It should also be observed that when X represents the same series of values the quantities ΣX and

ΣfX are algebraically the same, f being merely a symbol of operation showing that the X's were added in frequency groups.

The above calculation may be considerably shortened by selecting an assumed mean, A, near the middle of the series as origin and measuring the variable in units of class intervals. We shall take these two steps separately to show their individual effect upon shortening the calculation.

f	X'	fX'	f	$d = \frac{X'}{h}$	fd
2	20	40	2	2	4
2	10	20	2	1	2
$A = 74.5$ 4	0	0	4	0	0
5	− 10	− 50	5	− 1	− 5
2	− 20	− 40	2	− 2	− 4
15		$-30 = \Sigma fX'$	15		$-3 = \Sigma fd$

The X' series, or "reduced series," has been obtained by subtracting 74.5 from each of the X's in the preceding illustration. In order to obtain the mean from the calculation on the left it is necessary to add 74.5 to the mean of the X' values since each has been diminished by that amount, that is, $M = 74.5 + \frac{-30}{15} = 72.5$.

It will be noted that the X' values are replaced in the work at the right by d values, which are obtained by dividing the X's by the width of the class interval, h. In obtaining the mean of the whole series, therefore, the mean of the d's, or $\frac{\Sigma fd}{N}$, must be multiplied by h, before being added to the assumed mean, A. The work will then be $M = 74.5 + \frac{-3}{15} \times 10 = 72.5$.

Some students may understand the above method more clearly by the following algebraic proof. From the definition of X' we have

$$X' = X - A,$$

so that $$X = A + X'.$$

Furthermore, $$X' = dh.$$

Hence $$X = A + dh.$$

Summing over this expression (or adding member by member as many equations of this type as there are cases), we obtain

$$\Sigma X = \Sigma A + \Sigma dh.$$

Dividing by N, factoring out h (which is a constant throughout the summation), and noting that $\Sigma A = NA$, we obtain the required formula,

$$M = \frac{\Sigma X}{N} = A + \left(\frac{\Sigma fd}{N}\right)h. \quad \left\{\begin{matrix}\text{Mean for}\\ \text{distribution}\end{matrix}\right\} \quad (6)$$

The symbol of operation, f, has been inserted for convenience.

We shall next take a somewhat longer series in order to review the above procedure and note a check on the work. The following scores were made by a class in statistics on the Otis Self-Administering Test:

TABLE 13. ILLUSTRATING THE CALCULATION OF THE MEAN WITH CHECK

	Class	f	d	fd	Check d'	Check fd'	
	69.5–74.5	6	5	30	4	24	
	64.5–69.5	2	4	8	3	6	
	59.5–64.5	3	3	9	2	6	
	54.5–59.5	6	2	12	1	6	
	49.5–54.5	10	1	10	0	0	$A = 52$
$A = 47$	44.5–49.5	23	0	0	−1	−23	
	39.5–44.5	8	−1	−8	−2	−16	
	34.5–39.5	4	−2	−8	−3	−12	
	29.5–34.5	4	−3	−12	−4	−16	
	24.5–29.5	1	−4	−4	−5	−5	
		$N = 67$		$\Sigma fd = 37$		$\Sigma fd' = -30$	

$$M = 47 + \tfrac{37}{67} \times 5 = 47 + 2.76 = 49.76$$
$$M \text{ (Check)} = 52 - \tfrac{30}{67} \times 5 = 52 - 2.24 = 49.76$$

In the first of the above calculations for the mean the origin is taken at 47, opposite the largest frequency, 23, because it looks as if this would furnish a small Σfd. The d's are then tabulated 1, 2, 3, · · · and −1, −2, −3, · · · from this point and the fd products formed. The remainder of the calculation consists in substituting $\Sigma fd = 37$ in formula (6), where $A = 47$, $N = 67$, and $h = 5$.

The check on the right is made by selecting a new reference point or origin and repeating the calculation at least up to the quantity $\Sigma fd'$. If the new origin differs from the old by one class unit, $\Sigma fd'$ will differ from Σfd by N. This can be seen by inspection or shown as follows:

$$d = d' \pm 1,$$
$$\Sigma fd = \Sigma fd' \pm \Sigma f,$$
$$\therefore \Sigma fd = \Sigma fd' \pm N. \quad \{\text{Check on mean}\} \qquad (7)$$

In the above example, $d' = d - 1$ and $\Sigma fd'$ should equal $\Sigma fd - N$, which it does, checking the work to that stage in the calculation. The student is warned not to forget to multiply the quantity $\Sigma fd/N$ by the width of the interval h. Failure to do so is detected by carrying the check computation through to the final result. It is therefore desirable to use the complete check until the student is confident of the accuracy of his calculations.

3. Properties of the Mean

The arithmetical mean has several important properties which should be noted. First of all, it is rigorously defined in algebraic terms and is based directly on the actual values of all the items. This makes it possible to obtain a definite average for any quantitative series, and gives a result which is truly characteristic of the whole distribution.

The algebraic character of the mean makes possible the combination of averages from several series. Thus, if X_1, X_2, and X_3 denote the variables in three different groups of size N_1, N_2, and N_3, the three means will be $M_1 = \frac{\Sigma X_1}{N_1}$, $M_2 = \frac{\Sigma X_2}{N_2}$, and $M_3 = \frac{\Sigma X_3}{N_3}$. The mean of all three series is the sum of all the X's divided by the total number of items, or $M = \frac{\Sigma X_1 + \Sigma X_2 + \Sigma X_3}{N_1 + N_2 + N_3}$. This result may be obtained from the individual means by multiplying each mean by the size of its group and dividing the sum of these products by the total number of items,

that is, $M = \frac{N_1M_1 + N_2M_2 + N_3M_3}{N_1 + N_2 + N_3}$. This property might be of great advantage in combining norms from different localities, for it would be necessary to know only the means and the number of cases in each group.

It should be noted that the mean of several means is rarely the average of the items on which the separate means are based. This can be seen by a very simple example. Let $M_1 = \frac{6}{2} = 3$, $M_2 = \frac{20}{5} = 4$, and $M_3 = \frac{15}{3} = 5$. The mean of *all the items* is $\frac{6+20+15}{10}$, or 4.1, while the mean of *the three means* is $\frac{3+4+5}{3}$, or 4.0. Both results are entirely correct, but represent quite different things. The reason for the discrepancy may be seen by noting that if the three means are different, $\frac{N_1M_1 + N_2M_2 + N_3M_3}{N_1 + N_2 + N_3}$ will equal $\frac{M_1 + M_2 + M_3}{3}$ only when the three N's are equal. Since it is usually the mean of the *items* which is required, the averaging of averages should ordinarily be avoided.

Another property of the mean appearing from the definition is that every item, large or small, contributes its proportionate share to the result. This is regarded by some as a defect since extreme items appear to have an undue effect upon the mean. Against this objection it might be argued that if such extreme observations belong in the series, they should be permitted to contribute their full share.

The algebraic properties of the mean are of further importance in mathematical statistics where this average enters as a parameter in many formulas. It is probably as valuable in this respect as any other statistical constant.

A final, and in some respects the most valuable, property of the mean is its stability under fluctuations of sampling for ordinary distributions. If samples are drawn from a large body

of material and a number of means calculated, they will usually be closer to the mean of the whole material than if any other average had been employed. This property is often characterized as the *reliability* of the mean (see Chapter XIII).

4. Calculation of the Median

The median for ungrouped series has already been introduced in connection with the classifier described in Chapter II. It is the middlemost value of the variable when the observations are ranked in order of size, or the magnitude such that greater and smaller values occur with equal frequency. For an odd number of observations without ties in rank, it is clearly the magnitude of the middle observation. For an even number of cases any value between the two middle items will satisfy the above definition, but it is customary to take as the median the average of the two middle values. In case there are ties in rank near the middle of the series, a weighted average is sometimes used as illustrated by the following observations: 1, 3, 5, 9, 10, 12, 12, 12, 14, 14, 15, 16, 17, 21, 23, 25. The value halfway between 12 and 14 might serve as the median, but the weighted mean of the middle observations would seem to give a little more stable result. The median in this example is thus

$$\frac{3 \times 12 + 2 \times 14}{5} = 12.8.$$

When there are a sufficient number of observations to warrant the use of a frequency distribution, the above difficulties do not arise. The histogram of the Otis scores from Table 13 will illustrate the procedure in this case. Under this representation the frequencies are assumed to be spread evenly over the class intervals, the areas being exactly proportional to the number of items between any two class limits. The median is now to be regarded as the *value of the variable on either side of which half the frequencies lie.* The graphical solution amounts to determining the point on the scale the vertical through which bisects the

area under the histogram. It is thus only necessary to count in the frequencies from either end and interpolate across the interval containing the median.

From Fig. 22 it will be noted that half of the frequencies is 33.5, so that the problem is to determine the point above and below which 33.5 frequencies lie. Counting up from the lower end of the scale it is apparent that 17 frequencies lie below 44.5 and 40 frequencies lie below 49.5. The median therefore lies somewhere between these two values. The difference $33.5 - 17 = 16.5$ gives the number of frequencies beyond 44.5 necessary to reach the median. From the rectangles in the diagram it is apparent that $\frac{x}{5} = \frac{16.5}{23}$, so that the required distance, x, is $\frac{16.5}{23} \times 5 = 3.6$. The median is therefore $44.5 + 3.6 = 48.1$. The work may be checked by counting down from the upper end of the scale, giving median $= 49.5 - \frac{(33.5 - 27)}{23} \times 5 = 48.1$.

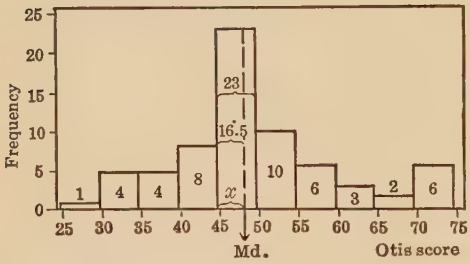


FIG. 22. Illustrating the median for the scores on the Otis Self-Administering Test

Using certain abbreviations, we may write two formulas for calculating the median in the case of the frequency distribution. The term *median interval* is used to designate the class interval which contains the median. Let

$u.l.$ and $l.l.$ = upper and lower limits of median interval, for example, 49.5 and 44.5 in Fig. 22,

f_{up} and f_{do} = total frequency up to and down to median interval, for example, 17 and 27,

f_{md} = frequency of median interval, for example, 23,

h = width of class interval, and

Md = median.

The formulas then become

$$Md = l.\,l. + \left(\frac{\frac{N}{2} - f_{up}}{f_{md}}\right) h \quad \left\{\begin{matrix}\text{Median for}\\ \text{distribution}\\ \text{counting up}\end{matrix}\right\} \quad (8a)$$

and

$$Md = u.\,l. - \left(\frac{\frac{N}{2} - f_{do}}{f_{md}}\right) h. \left\{\begin{matrix}\text{Median for dis-}\\ \text{tribution count-}\\ \text{ing down}\end{matrix}\right\} (8b)$$

If the student finds it easier to do the calculation by a series of steps, the following may be useful:

1. Divide the number of cases by 2, $\left(\frac{N}{2} = 33.5\right)$.

2. Determine by inspection the interval containing the median, $(44.5 - 49.5)$.

3. Count the frequencies up to the median interval, $(f_{up}=17)$.

4. Subtract this last result from $\frac{N}{2}$, $(33.5 - 17 = 16.5)$.

5. Multiply the last result by the width of the interval and divide by the number of frequencies in the median interval, $\left(\frac{16.5 \times 5}{23} = 3.6\right)$.

6. Add this quantity to the lower limit of the median interval, thus obtaining the required median, $(44.5 + 3.6 = 48.1)$.

A similar series of steps may be written out for the calculation when counting down from the upper end of the scale.

It has been noted that the median determines the point on the horizontal scale the vertical through which bisects the area of the histogram. The mean, on the other hand, is the point at which the histogram would balance. It is the center of gravity of the distribution. The *fd's* correspond to the moments in physics (force $\times$ distance), and the mean or center of gravity occurs where $\Sigma fd = 0$.

The table on page 88 shows the complete calculation of the mean and median for a longer series. It will be noted that the frequencies are given at central ages 45, 44, etc., or classes 44.5–45.5, 43.5–44.5, etc., since all ages were tabulated to the

nearest year. The data show the ages at which a group of college professors listed in Who's Who received their Ph.D. degrees. All these men had an A.B. but none an A.M. degree.

TABLE 14. ILLUSTRATING THE CALCULATION OF THE MEAN AND MEDIAN

Age Received Ph.D.	f	d	fd	Check d'	Check fd'	
45	3	16	48	17	51	
44	—	15	—	16	—	
43	3	14	42	15	45	
42	3	13	39	14	42	
41	1	12	12	13	13	
40	5	11	55	12	60	$M = 29 + \frac{-12}{400} \times 1 = 28.97$
39	190 9	10	90	11	99	
38	cases 5	9	45	10	50	$M = 28 + \frac{388}{400} \times 1 = 28.97$
37	above 5	8	40	9	45	
36	Md. 7	7	49	8	56	
35	Int. 7	6	42	7	49	
34	10	5	50	6	60	
33	13	4	52	5	65	
32	17	3	51	4	68	
31	29	2	58	3	87	$Md = 27.5 + \frac{17}{27} \times 1 = 28.13$
30	42	1	42	2	84	
29	31	0	—	1	31	
28	27	−1	−27	0	0	$Md = 28.5 - \frac{10}{27} \times 1 = 28.13$
27	37	−2	−74	−1	−37	
26	54	−3	−162	−2	−108	
25	183 38	−4	−152	−3	−114	
24	cases 29	−5	−145	−4	−116	
23	below 14	−6	−84	−5	−70	
22	Md. 7	−7	−49	−6	−42	
21	Int. 2	−8	−16	−7	−14	
20	2	−9	−18	−8	−16	
	400		−12		388	

5. PROPERTIES OF THE MEDIAN

The lack of rigor in the definition of the median for undistributed series has already been noted, and in this respect the mean is clearly superior. For large bodies of data, however, in which the use of the frequency distribution becomes imperative, no difficulties as to rigid definition are likely to arise.*

* Note that the median for grouped data becomes indeterminate when the frequency of the median interval is zero. This form of distribution, however, is very rare.

The median is based only indirectly upon all of the observations inasmuch as it is determined by their relative size. Whether or not this is an advantage over the mean depends upon the particular purpose for which the average is used. Under ordinary circumstances all items should contribute fully if included at all and the mean is therefore generally superior.

In combining the averages of several series the mean has a great advantage over the median. A simple combination of the separate means and totals as shown in section 3 will furnish the mean of the entire group of items, while in order to determine the grand median it is necessary to combine all of the separate distributions into one and calculate from this. As regards other algebraic properties the median is again inferior since it cannot be employed in connection with the formulas of higher statistical analysis.

The reliability of the median, or its stability under fluctuations of sampling, is in general less than that of the mean. Only for very peaked or leptokurtic distributions of the type illustrated in (5) of Fig. 21 is the median superior in this respect.*

The advantage thus far appears to be entirely in favor of the mean, but the median has at least two points of superiority. It is easier to calculate for both long and short series, and in the case of ungrouped data the middle item which furnishes the median can be uniquely identified and will remain the median item under any other form of measurement. Thus the height of the eleventh man in a group of twenty-one is typical of all in a very real sense, while the mean of the series will very probably not correspond to the height of any particular individual.

For the large bulk of test data the norms, or average scores for unselected groups, are given in the form of the medians. In using such tests and in making comparisons it is therefore necessary to use this form of average. For most problems, however, the mean is distinctly superior and should be used unless there is some very good reason to the contrary.

*G. U. Yule, Introduction to Statistics, p. 339. C. Griffin & Co., London.

6. The Crude Mode

The modal value of a variable is the value of the most frequent occurrence. Thus in Fig. 21 the modes are the abscissas corresponding to the highest points of the curves. For grouped series it is possible to obtain only a *crude mode,* which may be defined as the class value of the group with the largest frequency.

The crude mode is obviously unstable inasmuch as it will depend upon the fineness of classification used in grouping the data. By widening or narrowing the class interval, the mode may be made to shift very considerably up or down the scale. It is therefore to be used only for rough inspectional purposes. Its great advantage, of course, lies in the fact that it can be determined at a glance.

The following distribution shows two crude modes for the A.B. to A.M. spans of a group of college professors. The spans or years elapsing between degrees are again given at class values.

Years between A.B. and A.M. Degrees	f	f'
13	2	6
12	4	
11	3	6
10	3	
9	10	22
8	12	
7	8	23
6	15	
5	20	68
4	48	
3	144	238
2	94	
1	152	152
	515	515

In the first frequency distribution given by f, the crude modes appear at *one* and at *three* years. Grouping by two-year intervals brings a single mode at *two and one-half* years. The two crude modes are of greater practical interest in this example because they show that if a graduate student fails to get his master's degree in one year, he will very likely take three years

instead of two. This is probably due to the fact that he has had to leave the university for a time upon failure to complete the work in a year or that after taking the bachelor's degree he waited a year or two before working on the master's degree. In occasional problems of this sort the crude mode is of interest, but in general some other average should be used.

7. The Geometric Mean and Geometrical Series

Geometrical series, which were introduced in Chapter IV, will now be considered more generally and applied to some statistical problems. The geometric mean of a series of observations is the value obtained by finding the product of all the observations, and then obtaining the root of that product with an index equal to the number of items in the group. Thus the geometric mean of the values X_1, X_2, X_3, $\cdots$ X_N may be defined by the relation

$$G.M. = \sqrt[N]{X_1 \ X_2 \ X_3 \cdot \cdots \ X_N}, \quad \left\{ \begin{matrix} \text{Geometric} \\ \text{mean} \end{matrix} \right\} \quad (9)$$

or in terms of logarithms,

$$\log (G.M.) = \frac{1}{N} \Sigma \log (X), \quad \left\{ \begin{matrix} \text{Logarithmic} \\ \text{form of geo-} \\ \text{metric mean} \end{matrix} \right\} \quad (10)$$

the latter form furnishing the usual scheme of calculation.

It will be noted that the geometric mean becomes zero if any of the X's are zero, and may become imaginary if negative values occur. As shown in most texts in algebra, the geometric mean of a series will always be less than the arithmetic mean.

A *geometrical progression* has been defined in Chapter IV as a series of terms such that each term is the product of the preceding term by a constant factor called the *ratio*. In the geometrical series 8, 12, 18, 27, 40.5, this ratio is clearly 1.5, and the geometric mean of the whole series is

$$\begin{aligned} &\sqrt[5]{8 \times 12 \times 18 \times 27 \times 40.5} \\ &= \sqrt[5]{8 \times 8(1.5) \times 8(1.5)^2 \times 8(1.5)^3 \times 8\,(1.5)^4} \\ &= \sqrt[5]{8^5(1.5)^{10}} = 8 \times (1.5)^2, \end{aligned}$$

or

$$G.M. = 8 \times 2.25 = 18.$$

That this result should follow appears at once from the more general form of the geometrical series a, ar, ar^2, ar^3, ar^4, etc., the $G.M.$ of the first five terms being

$$\sqrt[5]{a \cdot ar \cdot ar^2 \cdot ar^3 \cdot ar^4} = \sqrt[5]{a^5 r^{10}} = ar^2,$$

or the middle term, which is 18 in the example on page 91.

The arithmetic mean of the above items is 21.1. In order to show the relationship between these two means, they have been plotted with the data in Fig. 23. As a general average of the five numbers in the series the arithmetic mean is quite adequate, but as a measure of the average item in such a *trend* the geometric mean only is correct. This is further illustrated by the means of the first four terms. The $A.M.$ is now 16.25 and the $G.M.$ mean 14.70, the latter being again a point on the smooth curve connecting the items in the series.

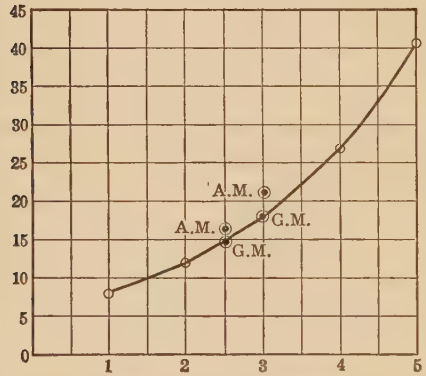


FIG. 23. Illustrating the geometric and arithmetic means for four and five items

The geometric mean is thus useful in determining averages in historical trends where the items form something like a geometrical series. The cost data in Table 15 furnish an example of this sort.

From the nature of a geometric series it is apparent that the ratio of each term to the one just preceding is equal to the constant ratio r. Applying this test to the cost data, a series of fairly equal ratios is found, showing that the original items form approximately a geometrical progression. In determining the expenditure at any point in this trend, the geometric mean should therefore be used. The total yearly costs have been taken at the mid-year points, so that if the cost from the

TABLE 15. COST DATA ILLUSTRATING THE USE OF THE GEOMETRIC MEAN

YEAR	EXPENDITURE FOR PUBLIC SCHOOLS IN THE UNITED STATES (IN MILLIONS)	RATIO OF EACH ITEM TO ONE ABOVE	THEORETICAL SERIES
1901–1902	227.5		230.0
1902–1903	238.3	1.047	246.9
1903–1904	252.8	1.061	265.1
1904–1905	273.2	1.081	284.8
1905–1906	291.6	1.067	305.8
1906–1907	307.8	1.056	328.5
1907–1908	336.9	1.095	352.8
1908–1909	371.3	1.102	378.9
1909–1910	401.4	1.081	406.9
1910–1911	426.3	1.062	437.0
1911–1912	446.7	1.048	469.3
1912–1913	482.9	1.081	504.1
1913–1914	521.5	1.080	541.4
1914–1915	555.1	1.064	581.4
1915–1916	605.5	1.091	624.5
1916–1917	640.7	1.058	670.7
1917–1918	702.2	1.096	720.3
1918–1919	763.7	1.088	773.6
	$A.M. = 435.9$	$A.M. = 1.0740$	
	$G.M. = 406.9$	$G.M. = 1.0739$	

middle of 1909 to the middle of 1910 were required it could be approximated by finding the geometric mean of 401.4 and 426.3, or $\sqrt{171{,}116.82}$, which is 413.7, or, since we are figuring in millions of dollars, $413,700,000.

For the entire series the $A.M.$ is 435.9 and the $G.M.$ 406.9, while for the set of accompanying ratios the arithmetic and geometric means agree at 1.074. The correct method of averaging such ratios is by the geometric mean, but in the above example there is very close agreement between the two averages because of the even nature of the series.

The theoretical series given in Table 15 was obtained by forming a geometric progression with $a = 406.9$ and $r = 1.074$, and extending it in both directions from the beginning of the year 1910. It may be noted that an error in the fourth place of

these numbers may occur because of the cumulative nature of the errors in the powers of r.

From the very good agreement between the observed and theoretical cost trends it is further apparent that the expenditure data form a good approximation to a geometrical series. From the years 1901 to 1918 the cost rose an average of

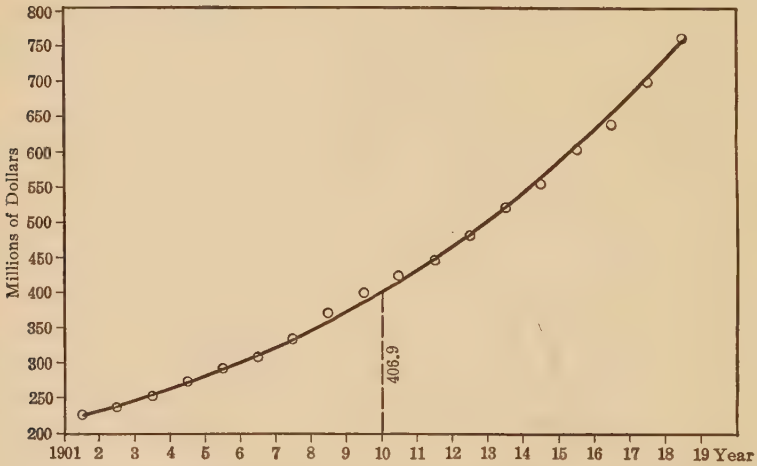


FIG. 24. Plot of the cost data and theoretical series from Table 15

7.4 per cent each year over the one just preceding. The character of this increase is the same as that of a sum of money out at compound interest.

The close agreement noted above might tempt one to extend the curve and predict future costs. Thus the expenditure in 1922 might have been forecasted as $(406.9)(1.074)^{12.5}$. This gives 993.2 as compared with an actual cost expenditure of 1,580.7 millions. The large discrepancy is due chiefly to the influence of post-war conditions upon the purchasing power of money. In making the prediction it was assumed that the same factors influencing expenditures from 1901 to 1918 would continue to operate in 1922. This assumption, as we have seen, is not valid.

8. The Harmonic Mean

The harmonic mean of a series of observations is the reciprocal of the arithmetic mean of their reciprocals, or if H (or $H.M.$) be the harmonic mean,

$$\frac{1}{H} = \frac{1}{N}\Sigma\left(\frac{1}{X}\right). \quad \{\text{Harmonic mean}\} \qquad (11)$$

For the series 8, 12, 18, 27, 40.5 the harmonic mean will thus be given by

$$\frac{1}{H} = \frac{1}{5}\left(\frac{1}{8} + \frac{1}{12} + \frac{1}{18} + \frac{1}{27} + \frac{1}{40.5}\right) = .06512.$$

Therefore $H = 15.4$.

The work can be done very readily using a table of reciprocals. The three averages for the above data may now be written

$$H.M. = 15.4,$$
$$G.M. = 18,$$
$$A.M. = 21.1,$$

which is the order of magnitude always found as shown by texts in algebra.

The harmonic mean may be illustrated by a supposititious problem. Let us assume that five pupils worked an hour on some problems, with the results set down in two forms as follows:

Problems Worked in an Hour		Minutes Required to work a Problem	
10		6	
8		7.5	
6		10	
4		15	
2		30	
M_r	6	M_t	13.7
H_r	4.38	H_t	10

If r and M_r denote the rate and arithmetic mean rate at which the problems were worked, and t and H_t represent the time and harmonic mean time in minutes required to work a problem, it is evident that

$$r = \frac{60}{t}$$

and $$\frac{\Sigma r}{N} = 60\,\frac{1}{N}\,\Sigma\left(\frac{1}{t}\right),$$

or $$M_r = 60\,\frac{1}{H_t}, \quad \left(6 = 60 \times \frac{1}{10}\right).$$

Similarly, $$M_t = 60\,\frac{1}{H_r}, \quad \left(13.7 = 60 \times \frac{1}{4.38}\right).$$

The arithmetic mean rates and mean times may therefore be obtained by determining the reciprocals of the harmonic mean times and rates and multiplying by the proper constant.

Two experimenters, for instance, might have recorded their results, one as rate and one in time, but by using the above relationships their averages could be made directly comparable. Thus in the above example the mean rate, 6, may be found by the arithmetic mean of the rates or by dividing 60 by the harmonic mean of the corresponding times.

In general, if the test results are recorded as rates, M_r should be employed; but if times are recorded, M_t should be used. The corresponding harmonic means are chiefly useful in making results comparable when necessary.

EXERCISES

1. Calculate the mean and median for each of the following frequency distributions:

(1)		(2)		(3)	
Class	f	Class	f	Class Value	f
94.5–99.5	1	36.5–38.5	1	10	1
89.5–94.5	2	34.5–36.5	–	9	2
84.5–89.5	3	32.5–34.5	3	8	5
79.5–84.5	5	30.5–32.5	4	7	10
74.5–79.5	7	28.5–30.5	10	6	12
69.5–74.5	6	26.5–28.5	4	5	10
64.5–69.5	4	24.5–26.5	3	4	4
59.5–64.5	–	22.5–24.5	2	3	3
54.5–59.5	1	20.5–22.5	2	2	–
				1	2
$M = 77.52$		$M = 28.81$		$M = 5.86$	
$Md = 77.00$		$Md = 29.20$		$Md = 5.96$	

(4)

Class	f
89.5–99.5	1
79.5–89.5	2
69.5–79.5	5
59.5–69.5	20
49.5–59.5	16
39.5–49.5	4
29.5–39.5	5
19.5–29.5	2
9.5–19.5	1

$M = 57.36$
$Md = 59.50$

(5)

Class Value	f
95	1
85	–
75	3
65	–
55	5
45	6
35	7
25	–
15	4
5	2

$M = 42.14$
$Md = 41.67$

(6)

Class	f
40.5–43.5	1
37.5–40.5	2
34.5–37.5	5
31.5–34.5	6
28.5–31.5	7
25.5–28.5	10
22.5–25.5	4
19.5–22.5	3
16.5–19.5	2

$M = 29.325$
$Md = 28.93$

(7)

Class	f
34.95–39.95	1
29.95–34.95	3
24.95–29.95	7
19.95–24.95	10
14.95–19.95	4
9.95–14.95	2
4.95–9.95	2

$M = 22.79$
$Md = 23.20$

(8)

Class	f
10.25–11.25	2
9.25–10.25	2
8.25–9.25	4
7.25–8.25	7
6.25–7.25	8
5.25–6.25	4
4.25–5.25	2
3.25–4.25	1

$M = 7.35$
$Md = 7.25$

(9)

Class Value	f
11.5	1
10.5	2
9.5	4
8.5	5
7.5	6
6.5	7
5.5	4
4.5	2

$M = 7.56$
$Md = 7.42$

2. Calculate the means and medians for the data of Exercise 1, Chapter II, using class intervals of 10 for the Otis and Terman tests, and an interval of 5 units for the Chicago test.

	Otis	Chicago	Terman
Mean	139.3	53.75	124.5
Median	140.6	53.64	124.5

Ans.

3. Verify the means and medians for the distributions of the Army Alpha Test given on pages 98 and 99. The class values were taken at 207.5, 202.5, etc., which makes the averages .5 larger than they would have been if the intervals had been given as 204.5–209.5, etc.

VARIABLES: ALPHA SCORE × SCHOOLING. GROUP I, II, III: WHITE DRAFT (NATIVE BORN)*

For men who took alpha only.

ALPHA SCORE	GRADES		HIGH SCHOOL			
	7	8	1	2	3	4
205–209	–	–	1	–	–	–
200–204	–	–	–	–	–	–
195–199	–	–	1	–	1	–
190–194	–	–	–	–	2	3
185–189	–	1	2	3	2	9
180–184	–	2	2	6	4	15
175–179	2	5	4	6	4	15
170–174	2	5	6	10	11	22
165–169	3	8	8	7	10	38
160–164	–	12	18	12	10	48
155–159	2	15	22	20	24	56
150–154	7	29	36	30	29	63
145–149	7	48	27	42	34	96
140–144	7	62	44	41	41	98
135–139	15	76	55	57	46	106
130–134	19	108	73	85	69	130
125–129	17	159	86	89	62	120
120–124	24	164	92	94	74	121
115–119	36	249	113	129	80	148
110–114	52	309	136	126	91	151
105–109	66	384	173	146	83	140
100–104	97	430	168	174	97	135
95–99	141	523	199	148	95	135
90–94	170	624	209	174	97	105
85–89	187	661	230	201	107	110
80–84	247	756	232	167	84	103
75–79	326	811	248	137	82	87
70–74	378	914	238	165	78	81
65–69	385	957	225	131	57	70
60–64	499	989	246	146	64	57
55–59	594	1,057	178	114	38	33
50–54	611	996	161	95	21	32
45–49	650	937	143	73	24	18
40–44	660	845	107	55	24	21
35–39	638	706	88	49	27	14
30–34	636	642	80	36	13	16
25–29	511	461	45	31	12	11
20–24	380	281	27	12	8	7
15–19	231	189	10	16	4	8
10–14	54	59	1	7	4	–
5–9	44	34	1	2	1	1
0–4	3	10	1	2	–	–
Total	7,701	14,518	3,736	2,838	1,614	2,423
M	53.874	68.287	83.842	90.366	98.823	109.881
Md	50.356	65.277	81.487	89.502	98.263	110.911

* Data from Memoirs of National Academy of Sciences, Vol. XV, p. 748. *Ans.*

VARIABLES: ALPHA SCORE × SCHOOLING. GROUP I, II, III: WHITE DRAFT (NATIVE BORN)* (CONTINUED)

For men who took alpha only.

ALPHA SCORE	COLLEGE					
	1	2	3	4	5	6
205–209	–	–	–	–	–	–
200–204	–	–	1	2	–	–
195–199	2	1	1	5	1	1
190–194	1	2	2	5	–	1
185–189	4	4	4	17	2	1
180–184	2	3	2	19	1	–
175–179	9	8	6	23	2	1
170–174	9	13	12	36	5	1
165–169	16	14	14	38	1	2
160–164	25	18	17	44	4	3
155–159	22	29	28	39	2	1
150–154	29	34	31	47	5	1
145–149	33	31	28	53	6	2
140–144	26	26	18	41	3	–
135–139	51	37	21	29	3	2
130–134	42	50	27	29	1	1
125–129	62	38	29	35	4	1
120–124	48	43	20	34	2	1
115–119	44	58	28	42	1	–
110–114	65	48	28	28	–	1
105–109	52	36	23	21	6	2
100–104	44	33	22	17	1	–
95–99	58	25	24	26	2	–
90–94	47	51	25	17	–	–
85–89	55	35	21	10	–	–
80–84	41	37	17	11	1	–
75–79	39	29	9	6	3	2
70–74	45	22	12	4	1	2
65–69	29	29	13	13	–	–
60–64	41	17	8	2	–	–
55–59	38	16	11	2	–	–
50–54	16	12	6	3	–	–
45–49	23	11	8	3	2	–
40–44	11	7	3	–	1	–
35–39	17	3	1	1	–	–
30–34	3	2	2	2	–	–
25–29	2	3	–	2	–	–
20–24	1	4	1	–	–	–
15–19	4	–	–	–	–	–
10–14	–	–	–	1	–	–
5–9	–	–	–	–	–	–
0–4	–	–	–	–	–	–
Total	1,056	829	523	707	60	26
M	105.559	112.168	118.877	136.581	134.417	140.0
Md	106.346	114.427	119.911	141.890	143.333	147.5

Ans.

* Data from Memoirs of National Academy of Sciences, Vol. XV, p. 748.

4. Calculate the geometric mean for the following cost data:

Year	Total Expenditure for Schools in United States Relative to 1914
1914	100
1916	115
1918	138
1920	187
1922	285

($G.M. = 153$)

Taking $a = 100$ and $ar^4 = 285$, compute r and construct a geometrical series of five terms ($r = 1.30$). Compare with the data. Should you conclude that the cost increased in geometrical progression during this period?

CHAPTER VII

MEASURES OF DISPERSION

1. Introductory

The dispersion of a series of observations is the degree of scatter, or the extent to which the items are spread out along the scale from some average value. It is important to have measures of such variability for several reasons, one of them arising from its relation to the reliability of the average.

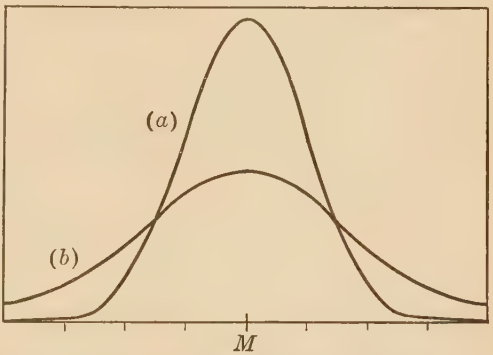


Fig. 25. Illustrating difference in dispersion for two series with the same mean

In Fig. 25 two distributions with the same number of cases and the same average are shown. In the case of curve (*a*) the observations cluster closely around the mean, while in curve (*b*) they are spread out much more along the scale. It is therefore apparent that the average of the first distribution is more representative of the whole series, more typical of all the observations, and for the same number of items to be regarded as the more reliable. In comparing two or more averages it is necessary to have some measure of their respective reliabilities, and for this purpose a numerical representation of the dispersion of the series is first required. Appropriate reliability formulas for averages and other statistical quantities will be found in Chapter XIII.

Another need for a measure of dispersion arises when a general measure of homogeneity is required. In teaching, for example, it is well known that when the students in a class differ widely in previous training and mental characteristics, instruction becomes a very difficult problem. The measurement of the abilities involved and the quantitative appraisal of the variation in different groups make it possible to approach such a problem scientifically. Closely related to this problem is the question whether or not uniform instruction tends to bring a class up to a common level of attainment, or brings about a still further differentiation in ability. These questions can be answered best by making use of some measure of dispersion.

Other uses of group variability appear in connection with problems in the overlapping of pupil abilities and in the standardization of tests, as a parameter in higher statistical analysis, and as a common unit of measure in comparing performances on unlike scales. This last use will be discussed in section 7.

2. Mean Deviation

One of the simplest measures of dispersion is the mean deviation or variation of the observations about some central tendency such as the arithmetic mean or median. The computation will first be illustrated by a short ungrouped series. If x denotes the variation of an observation X from the mean M, then $x = X - M$ and the original values and deviations for five scores may be set down as follows:

X	x
21	+ 4.4
19	+ 2.4
17	+ 0.4
14	− 2.6
12	− 4.6
$M = 16.6$	$14.4 = \Sigma \lvert x \rvert$

It will be observed that the algebraic sum of the deviations x is zero in the above problem. This may be generally shown

by noting that if $X' = X - A$, then $\frac{\Sigma X'}{N} = M - A$, so that the sum of the deviations X' vanishes when $M = A$. In securing a measure of average variation it is therefore necessary to eliminate the algebraic signs in some way. The mean deviation is secured by adding the absolute values of the deviations (disregarding sign) and dividing by their number, or in symbols,

$$M.D. = \frac{\Sigma |x|}{N}. \quad \{\text{Mean deviation}\} \quad (12)$$

In the illustrative example,

$$M.D. = \frac{14.4}{5} = 2.88.$$

This simple process becomes lengthy if the mean and deviations are written to several decimal places, and for this reason a shorter method will next be introduced. The procedure is illustrated by Fig. 26. The above five scores are represented by the horizontal bars, and the deviations from the mean by the hatched and dotted portions. Since the total negative deviation is equal to the total positive deviation, the deviation for the entire series may be obtained by determining the total negative deviation and multiplying the result by 2. Furthermore, the negative deviation may be found by subtracting from the sum of the segments AC and DF the sum of the original observations represented by AB and DE.

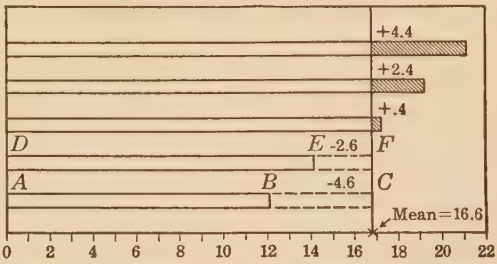


FIG. 26. Illustrating deviations from the mean for five scores

The complete procedure may then be described as follows:

1. Arrange the observations in order of size.
2. Compute the mean, (16.6).
3. Count the items smaller than the mean and multiply their number by the mean, ($2 \times 16.6 = 33.2$).

4. Subtract from this result the sum of the items smaller than the mean, $(33.2 - 26 = 7.2)$. This is the total negative deviation.

5. Multiply this last result by two, $(2 \times 7.2 = 14.4)$, and divide by N to determine $M.D.$, $(14.4/5 = 2.88)$.

The work through step 4 may be checked by adding the items larger than the mean and subtracting from this sum the product of the mean by the number of greater items.

It is evident graphically that if the same quantity is added to or subtracted from each item the deviations remain unchanged. This may also be shown algebraically. If A is the quantity subtracted we may write

$$X' = X - A,$$

$$\frac{\Sigma X'}{N} = \frac{\Sigma X}{N} - A,$$

so that $$M' = M - A,$$

and $$x' = x.$$

This simplification of the items is occasionally useful in further shortening the calculation. The whole procedure is illustrated by another series as shown in Table 16.

TABLE 16. SHOWING THE CALCULATION OF MEAN DEVIATION FOR AN UNGROUPED SERIES

X	$X' = X - 110$	
123	13	$6 \times 5.182 = 31.09$
120	10	$-\ 9.$
119	9	22.09
119	9	$\times\ 2$
117	7	44.18
114	4	
112	2	$M.D. = \frac{44.18}{11} = 4.02$
112	2	*Check:*
111	1	$48.$
110	0	$5 \times 5.182 = 25.91$
110	0	22.09
	$5.182 = M'$	

5 items larger than M': Sum of items larger than $M' = 48$

6 items smaller than M': Sum of items smaller than $M' = 9$

In determining mean deviation it is theoretically better to take the deviations from the median instead of from the mean because, as can be readily demonstrated, the total variation is less about the median. The above short method could not have been used, however, if the median had been employed, since the sum of the deviations about this average is not zero. Furthermore, for longer series it makes very little difference numerically which average is selected. The mean may, therefore, be used in ordinary practice.

In the case of the frequency distribution the same method may be used as for ungrouped items, the values of the observations being taken at the mid-points of the intervals, that is, at class values. The work is illustrated with the following problem:

Class	f	d	fd	
90–100⁻	1	3	3	$8 \times 63 = 504$
80–90⁻	2	2	4	370
70–80⁻	4 (12)	1	4 (11)	134
60–70⁻	5	0	0	2
50–60⁻	4	−1	−4	268
40–50⁻	3	−2	−6	$M.D. = \frac{268}{20} = 13.4$
30–40⁻	− (8)	−3	(−15)	*Check:*
20–30⁻	−	−4		890
10–20⁻	1	−5	−5	$12 \times 63 = 756$
	20		−4	134

The class values in this example are 15, 25, 35, etc., so that the mean is $65 - \frac{4}{20} \times 10$, or 63, by formula (6). There are 8 frequencies below 63, and 12 above, since the 5 in the interval 60–70⁻ comes at 65. The product of 8 and 63 gives a result equal to the sum of the items smaller than the mean plus their deviations from the mean. Next, the sum of the products of the smaller class values by their corresponding frequencies is $55 \times 4 + 45 \times 3 + 15 \times 1 = 370$. Subtracting this last result from 504 furnishes the total negative deviation 134. Multiplying this result by 2 to obtain the total positive and negative deviation, and dividing by 20 gives $M.D. = 13.4$.

By making use of certain abbreviations, a formula for mean deviation may now be set up. Let

A_m = the class value of the interval in which M lies,

N_a and N_b = the number of observations above and below M,

T_a and T_b = the sums of the observations above and below M,

$\Sigma \mid fd \mid_a$ and $\Sigma \mid fd \mid_b$ = absolute values of the parts of Σfd above and below A_m, and

h = the width of the class interval.

The steps in the calculation on page 105 may now be combined so as to give the checking formula

$$M.D. = \frac{2(T_a - N_a M)}{N} = \frac{2(N_b M - T_b)}{N}$$

or

$$M.D. = \frac{T_a - T_b - M(N_a - N_b)}{N}. \qquad (13)$$

It then remains to find T_a and T_b for the frequency distribution. These are clearly given by

$$T_a = N_a A_m + (\Sigma \mid fd \mid_a)h$$

and

$$T_b = N_b A_m - (\Sigma \mid fd \mid_b)h.$$

Substituting these values in equation (13) and noting that

$$\Sigma \mid fd \mid_a + \Sigma \mid fd \mid_b = \Sigma \mid fd \mid,$$

we have

$$M.D. = \frac{(\Sigma \mid fd \mid)h + (A_m - M)(N_a - N_b)}{N}, \left\{ \begin{matrix} \text{Mean deviation} \\ \text{for frequency} \\ \text{distribution} \end{matrix} \right\} \qquad (14)$$

which is the desired result.

Applying formula (14) to the problem on page 105, we find that

$$M.D. = \frac{26 \times 10 + (65 - 63)(12 - 8)}{20} = \frac{268}{20} = 13.4, \text{ as before.}$$

In order to fix the method of calculation and to warn the student of the difficulty which arises when A is not taken in the interval in which M lies, another model problem is next given with complete computations.

TABLE 17. ILLUSTRATING THE CORRECT AND INCORRECT CALCULATION OF MEAN DEVIATION FOR A DISTRIBUTION USING FORMULA (14)

CORRECT METHOD					INCORRECT METHOD		
Class Value	f		d	fd	d'	fd'	
97.5	22		3	66	5	110	
92.5	68	$N_a = 169$	2	136	4	272	
87.5	51		1	51	3	153	638
82.5	28		0	—	2	56	
77.5	47		−1	−47	1	47	
72.5	33		−2	−66	0	0	
67.5	21		−3	−63	−1	−21	
62.5	9	$N_b = 120$	−4	−36	−2	−18	
57.5	6		−5	−30	−3	−18	−76
52.5	2		−6	−12	−4	−8	
47.5	1		−7	−7	−5	−5	
42.5	1		−8	−8	−6	−6	
	$N = 289$			$\Sigma fd = -16$		$\Sigma fd' = 562$	

$N_a - N_b = 49$ $\qquad \Sigma \mid fd \mid = 522$

$$M = 82.5 - \frac{16 \times 5}{289} = 82.223$$

$$A_m - M = .277$$

$$M.D. = \frac{522 \times 5 + (.277)49}{289}$$

$$= \frac{2623.573}{289} = 9.078$$

$\Sigma \mid fd' \mid = 714$

$$M.D. \neq \frac{714 \times 5 - 9.723 \times 49}{289}$$

$$\neq 10.70$$

The work on the left is correct, while that on the right with origin at 72.5 is quite wrong. In case it is found that A does not lie in the interval containing the mean, this should be adjusted at once, using the previous results as a check on the mean. The reason for the incorrectness of the method on the right may be shown by noting that the expressions for T_a and T_b on page 106 give incorrect results in this case. The complete proof is left as an exercise for the student.

3. The Standard Deviation

In order to introduce the next measure of dispersion we may return to the short series shown at the beginning of the preceding section. A measure of average deviation was there found by adding the deviates from the mean regardless of sign. By the present method the algebraic signs are eliminated by squaring the deviations from the mean.

X	x	x^2
21	+ 4.4	19.36
19	+ 2.4	5.76
17	+ 0.4	.16
14	− 2.6	6.76
12	− 4.6	21.16
$M = 16.6$		$\Sigma x^2 = 53.20$

$$S.D. = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{53.2}{5}}$$
$$= \sqrt{10.64} = 3.26$$

The quantity $\frac{\Sigma x^2}{N}$ might now be used as a measure of mean square dispersion, but it has been found much more convenient and theoretically desirable to take the square root of this average. The standard deviation is therefore defined as

$$S.D. = \sqrt{\frac{\Sigma x^2}{N}}. \quad \left\{ \begin{matrix} \text{Standard deviation,} \\ \text{original form} \end{matrix} \right\} \quad (15)$$

The method of calculation for ungrouped series is comparatively simple, but in order to obviate the squaring of decimals a short cut is usually employed.

It has been shown in section 2 that

$$x = X - M = x' = X' - M'.$$

Therefore $$x^2 = (X')^2 - 2\,X'M' + (M')^2$$

and $$\Sigma x^2 = \Sigma (X')^2 - 2\,M'(\Sigma X') + N(M')^2.$$

But since $$NM' = \Sigma X',$$

we may write $$\frac{\Sigma x^2}{N} = \frac{\Sigma (X')^2}{N} - (M')^2,$$

or $$S.D. = \sqrt{\frac{\Sigma (X')^2}{N} - (M')^2}. \quad \left\{ \begin{matrix} \text{Standard deviation for reduced series} \end{matrix} \right\} \quad (16)$$

Applying this formula to the above problem we have

X	$X' = X - 12$	$(X')^2$	
21	9	81	$S.D. = \sqrt{\frac{159}{5} - 21.16}$
19	7	49	$= \sqrt{10.64}$
17	5	25	$= 3.26,$
14	2	4	as before.
12	0	0	
	$M' = 4.6$	$\Sigma(X')^2 = 159$	

For the frequency distribution the same method is employed. Since $X' = dh$ and $M' = (\Sigma fd)h/N$ (see Chapter VI, section 2) the formula becomes

$$S.D. = \left(\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}\right)h, \left\{\begin{matrix}\text{Standard de-}\\ \text{viation for}\\ \text{distribution}\end{matrix}\right\} \quad (17)$$

the calculation being carried through to the last step in class units when the result is then multiplied by the width of the class interval h.

The work will be illustrated by the Otis test data from Table 13. It is necessary to calculate only one column of items in

TABLE 18. ILLUSTRATING THE COMPUTATION OF STANDARD DEVIATION FOR A DISTRIBUTION WITH CHECK

CLASS INTERVAL	f	d	fd	fd^2	d'	fd'	$f(d')^2$
69.5–74.5	6	5	30	150	4	24	96
64.5–69.5	2	4	8	32	3	6	18
59.5–64.5	3	3	9	27	2	6	12
54.5–59.5	6	2	12	24	1	6	6
49.5–54.5	10	1	10	10	0	—	—
44.5–49.5	23	0	—	—	−1	−23	23
39.5–44.5	8	−1	−8	8	−2	−16	32
34.5–39.5	4	−2	−8	16	−3	−12	36
29.5–34.5	4	−3	−12	36	−4	−16	64
24.5–29.5	1	−4	−4	16	−5	−5	25
	67		37	319		−30	312
			Σfd	Σfd^2		$\Sigma fd'$	$\Sigma f(d')^2$

$$\sigma^* = \left[\sqrt{\tfrac{319}{67} - \left(\tfrac{37}{67}\right)^2}\right] \times 5 = \left[\sqrt{4.7612 - .3049}\right] \times 5 = 2.11 \times 5 = 10.55.$$

$$\textit{Check: } \sigma = \left[\sqrt{\tfrac{312}{67} - \left(\tfrac{30}{67}\right)^2}\right] \times 5 = \left[\sqrt{4.6567 - .2005}\right] \times 5 = 2.11 \times 5 = 10.55.$$

* The standard deviation is frequently symbolized by the small Greek letter σ.

addition to the computation for the mean. The quantities fd^2 are obtained by multiplying each value of d by the corresponding fd products. These may be checked by multiplying f by d^2. Thus, $5 \times 30 = 150$, $4 \times 8 = 32$, etc., or $6 \times 25 = 150$, $2 \times 16 = 32$, etc.

A more complete check may be made by choosing a new origin as in the calculation for the mean. If

$$d = d' \pm 1,$$

$$d^2 = (d')^2 \pm 2d' + 1,$$

and

$$\Sigma fd^2 = \Sigma f(d')^2 \pm 2\,\Sigma fd' + N. \quad \left\{\begin{matrix}\text{Check on}\\ \text{standard}\\ \text{deviation}\end{matrix}\right\} \quad (18)$$

In the above problem $d = d' + 1$, so that Σfd^2 should equal $\Sigma f(d')^2 + 2\,\Sigma fd' + N$. Since $319 = 312 + 2(-30) + 67$, the work is checked to this stage in the calculation. The remainder of the computation consists in substituting the appropriate values in formula (17) as shown in the work under the model problem in Table 18. It will be noted that it is desirable to carry the work under the radical to four decimal places if the answer be required to two.

Before comparing the above two measures of dispersion and noting their uses, another measure of variability will be introduced. This is known as the semi-inter-quartile range, or more briefly, as the quartile deviation.

4. The Quartile Deviation

This measure of variability is defined as half the range of the middle 50 per cent of the observations when arranged in order of size or in a frequency distribution. It is only necessary to determine two values, Q_1 and Q_3, below and above which one quarter of the measures lie. The range Q_1 to Q_3 then includes the middle half of the observations and the semi-inter-quartile range is defined by the expression

$$Q = \frac{Q_3 - Q_1}{2}. \quad \{\text{Quartile deviation}\} \quad (19)$$

In the case of ungrouped material the work may often be done by inspection as shown in the accompanying table of total state and local per capita expenditures in southern states. Maryland is the median state with an expenditure of $6.11 per capita.

TABLE 19. PER CAPITA EXPENDITURES FOR EDUCATION IN SEVENTEEN SOUTHERN STATES

STATE	PER CAPITA EXPENDITURE FOR EDUCATION IN 1900	
Oklahoma	$11.94	
District of Columbia	10.68	
Delaware	9.02	
West Virginia	8.75	$Q_3 = \$8.58$
Texas	8.41	
Florida	7.72	
Louisiana	6.65	
Virginia	6.61	
Maryland	6.11	$Q_2 = Md = \$6.11$
North Carolina	5.44	
Tennessee	4.96	
South Carolina	4.63	
Arkansas	4.62	$Q_1 = \$4.59$
Alabama	4.55	
Georgia	4.55	
Mississippi	4.54	
Kentucky	4.36	

$$Q = \frac{\$8.58 - \$4.59^-}{2} = \frac{\$3.99^+}{2} = \$2.00.$$

The value for Q_3 is taken halfway between the expenditures for Texas and West Virginia, or at $8.58, and similarly for Q_1, which is $4.59. Q is then half the difference between these two results, or $2.00. It will be noted that nine cases lie between Q_1 and Q_3, and that this is more than half of the total number of items, which is seventeen. For so few items, however, it is hardly worth while to strive for a more accurate result, the purpose of the table being to furnish only rough comparisons.

The differences $Q_3 - Q_2 = \$2.47$ and $Q_2 - Q_1 = \$1.52$ are not equal to Q, because of the lack of symmetry in the series, but their sum is of course equal to $2\,Q$. With this limitation in mind

the value of Q may be said to furnish approximately the magnitude which, when laid off on both sides of the median, will include the middle half of the items.

When the data are in a distribution, the values for Q_1 and Q_3 are computed in the same way as the median, the only difference being that one quarter instead of one half of the observations are counted in from either end. The procedure may be illustrated for the following distribution of intelligence quotients. These data are taken from a survey made in several counties in Illinois, the results of the study being as yet unpublished.

TABLE 20. ILLUSTRATING THE COMPUTATION OF QUARTILE DEVIATION FOR A DISTRIBUTION

I.Q.	f	
150–160	2	
140–150	12	
130–140	36	f_{do}
120–130	103	
110–120	318	
100–110	799	$= f_3$
90–100	1074	
80–90	1059	
70–80	868	$= f_1$
60–70	366	
50–60	163	f_{up}
40–50	25	
30–40	9	
	4834	

$$Q_3 = 110 - \frac{1208.5 - 471}{799} \times 10 = 100.770$$

$$Q_1 = 70 + \frac{1208.5 - 563}{868} \times 10 = 77.437$$

$$Q_3 - Q_1 = 23.333$$

$$\therefore Q = 11.67$$

Q_3 and Q_1 may be computed most readily from formulas similar to those used for the median. If one quarter of the cases be counted in from either end of the distribution the formulas become

$$Q_3 = u.l. - \frac{\frac{N}{4} - f_{do}}{f_3} \times h \qquad (20a)$$

$$\left\{\begin{matrix}\text{Quartiles} \\ \text{for distribution}\end{matrix}\right\}$$

and

$$Q_1 = l.l. + \frac{\frac{N}{4} - f_{up}}{f_1} \times h, \qquad (20b)$$

where f_1 and f_3 are the frequencies of the intervals where Q_1 and Q_3 occur and the other symbols are used as in the formulas for the median. The calculation is shown in full at the right of the distribution. A check may be made by counting in three quarters of the way from either end of the distribution and using similar formulas.

5. Comparison of Measures of Dispersion

In order to bring together the quantitative methods discussed thus far, all the simple averages and measures of dispersion have been computed for the above distribution and located graphically on a histogram. The student should work out and verify the following results:

Mean = 89.28	$M.D.$ = 13.65
Median = 89.31	$S.D.$ = 16.86
Crude mode = 95.00	Q = 11.67

The close agreement of the mean and median would seem to indicate a high degree of symmetry in the distribution, but contrary to expectation the data do not furnish a good example of a normal probability curve as will be shown in Chapter XIII.

As illustrated by Fig. 27 a range of 2 Q includes the middle 50 per cent of the observations, a range of 2 $M.D.$ (from the mean) somewhat more than half of the cases, while a range of 2 $S.D.$ includes about two thirds of the items. Furthermore, the ratio of Q to $S.D.$ is approximately .69, while the ratio of $M.D.$ to $S.D.$ is .81. These are typical of the results found with fairly symmetrical distributions. For the normal probability curve these two ratios are .6745 and .7979 respectively (Chapter XII).

By laying off the standard deviation three times to the left and to the right of the mean, a range of 6 $S.D.$ from 38.70 to 139.86 is obtained. By referring to Table 20 for the frequencies it will be noted that about 4811 cases, or 99 per cent of all the observations, lie within this range. For distributions of this

type, then, deviations greater than 3 *S. D.* from the mean occur very infrequently. Similarly, a range of $7\frac{1}{2}$ *M. D.* will extend from 88.09 to 140.47, while a range of 9 *Q* (laid off from the median) runs from 86.80 to 141.82. Within all three of the above ranges, therefore, more than 99 per cent of the cases will ordinarily occur.

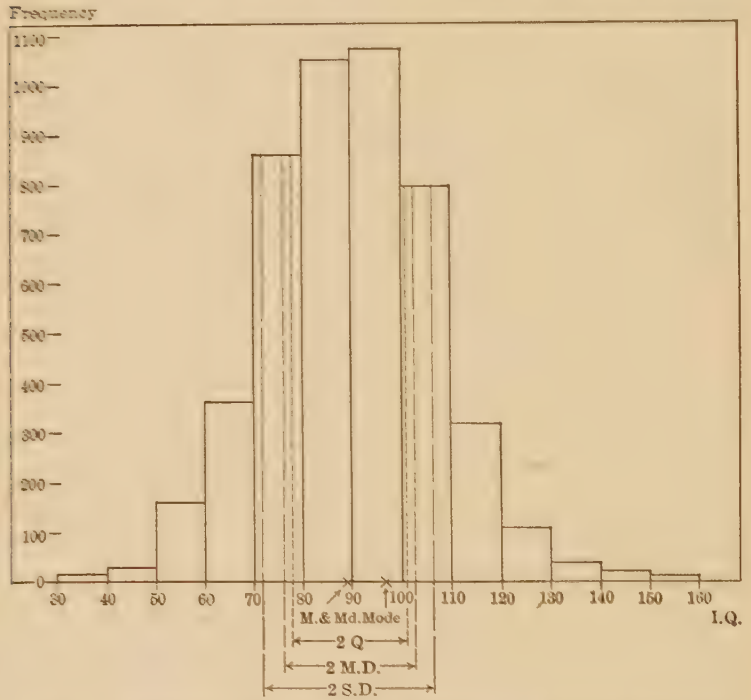


FIG. 27. Illustrating the comparative magnitude of several measures of dispersion

As regards clear definition there is little choice between the three measures of dispersion when the data are arranged in a frequency distribution. For undistributed series, however, the quartile deviation has the same defects as the median. As illustrated in Table 19, it is sometimes necessary to take the average of two neighboring values for Q_1 or Q_3.

The algebraic properties of the standard deviation make it the most useful for combining the results from several series, and in connection with other statistical formulas. Thus, if two series of size N_1 and N_2 have a total population of $N = N_1 + N_2$, with means M_1 and M_2, and standard deviations σ_1 and σ_2, it is possible from these values to find the mean M, and the standard deviation σ, of the whole group.

It has already been shown in Chapter VI that

$$M = \frac{N_1 M_1 + N_2 M_2}{N}.$$

The standard deviation of the total series may also be found. From the proof in section 3 it is apparent that if the assumed mean A be taken equal to M for both series,

$$M_1 - M = C_1$$

and

$$M_2 - M = C_2.$$

The mean square variations of the component series about M are, by equation (16), $\frac{\Sigma(X'_1)^2}{N_1} = \sigma_1^2 + C_1^2$ and $\frac{\Sigma(X'_2)^2}{N_2} = \sigma_2^2 + C_2^2$, respectively. The total square variation, or ΣX^2, of both groups about M is therefore

$$\Sigma(X'_1)^2 + \Sigma(X'_2)^2 = N_1(\sigma_1^2 + C_1^2) + N_2(\sigma_2^2 + C_2^2),$$

or

$$N\sigma^2 = N_1(\sigma_1^2 + C_1^2) + N_2(\sigma_2^2 + C_2^2), \qquad (21)$$

and in case the component series are of equal size we may write

$$\sigma^2 = \tfrac{1}{2}(\sigma_1^2 + \sigma_2^2).$$

The quartile deviation is probably the easiest measure of variability to compute, the mean deviation next, and the standard deviation most laborious of all. Simplicity of calculation, however, should rarely determine which measure of dispersion to employ since other properties are much more important.

The standard deviation is, in general, less affected by fluctuations in sampling than Q or $M.D.$, and for this reason alone is

preferable to the others. It is sometimes argued that the presence of a few extremely large or small observations may affect the standard deviation unduly, but if such items are truly a part of the distribution this objection is overruled.

In view of all of the above properties the standard deviation is the best measure of variability to employ for the fairly symmetrical distributions ordinarily found with educational or psychological data. A fairly safe rule with such material is to use the mean and the standard deviation whenever the data warrant careful treatment, reserving the median and Q for rough work with small samples.

6. The Coefficient of Variation

The measures of variability discussed thus far have two properties that are at once apparent.

1. They are expressed in the units of the variable so that direct comparisons of dispersion can be made only between series on the same scale.

2. They depend upon the size of the deviations from some central tendency, but are quite independent of the magnitude of the average itself.

A measure of variability which is independent of the scale units and which takes into account the size of the deviations relative to the mean may be expressed in the form $\sqrt{\Sigma\left(\frac{x}{M}\right)^2/N}$, which reduces at once to $\frac{\sigma}{M}$. Professor Pearson has called this quantity (when multiplied by 100 for convenience) the *coefficient of variation,* or percentage ratio of the standard deviation to the arithmetic mean. Denoting this new measure of variation by V, we have

$$V = \frac{100\,\sigma}{M}. \quad \{\text{Coefficient of variation}\} \quad (22)$$

It should be noted that while σ is the standard deviation of X, V is the standard deviation of $100\,X/M$. The student who

has difficulty in visualizing the significance of the coefficient of variation may thus regard it as the dispersion found when all of the observations (or deviations from the mean) have been made comparable by dividing each by $M/100$.

Direct comparisons of measures of absolute variability such as standard deviation, and relative variability as given by the coefficient of variation, often lead to confusion. Both are root mean square measures of variability, but of quite different things as shown above.

A simple example may illustrate this point:

$$M_1 = 20 \text{ problems}, \quad \sigma_1 = 4 \text{ problems}, \quad \therefore V_1 = 20$$
$$M_2 = 40 \text{ problems}, \quad \sigma_2 = 4 \text{ problems}, \quad \therefore V_2 = 10$$

These two series are equally variable as to absolute dispersion, but the relative variability in the first group is twice that in the second. Both measures are entirely correct, although it has been argued by Franzen* that the coefficient of variation should not be used with such material because of the arbitrary nature of the zero point on educational tests and scales. This amounts to objecting to the coefficient of variation because the size of the mean is arbitrary, but on the same grounds we should object to the use of the mean itself.

The chief use of the coefficient of variation is in comparing the dispersion of series where the means differ considerably in size and where the variation relative to the mean is therefore important.

The following distributions (p. 118) give the per capita state and local expenditures of forty-nine states (including the District of Columbia) for elementary and secondary education, and for all purposes in 1920.

If the standard deviation had been employed in comparing the variability of these two groups, it would have appeared that there is much more uniformity among the states in educational expenditure than in total expenditures. Using the coefficient

*Raymond Franzen, "Statistical Issues," *Journal of Educational Psychology*, September, 1924, p. 381.

TABLE 21. PER CAPITA EXPENDITURE FOR EDUCATION IN 49 STATES *

EXPENDITURE	f
$21–$22.99	1
$19–$20.99	2
$17–$18.99	3
$15–$16.99	6
$13–$14.99	4
$11–$12.99	6
$9–$10.99	9
$7–$8.99	7
$5–$6.99	3
$3–$4.99	8
Total	49

$M = \$10.94$
$\sigma = \$4.87$
$V = 44.5$

TABLE 22. PER CAPITA EXPENDITURE FOR ALL PURPOSES IN 49 STATES *

EXPENDITURE	f
$100–$109.99	1
$90–$99.99	–
$80–$89.99	1
$70–$79.99	1
$60–$69.99	8
$50–$59.99	7
$40–$49.99	6
$30–$39.99	12
$20–$29.99	6
$10–$19.99	7
Total	49

$M = \$43.16$
$\sigma = \$20.07$
$V = 46.5$

of variation, however, we find little difference in relative dispersion. While both results are correct for some purposes, the latter gives the better measure of the relative extent to which these two types of expenditure have become stabilized. A variation of a dollar in the first group is comparable with a variation not of one but of about four dollars in the second series. For such problems the relative rather than the absolute dispersion should be used to show the degree of uniformity in expenditure.

7. COMPARABLE MEASUREMENTS

One of the most important uses of variability is in furnishing units for the comparison of measurements on unlike scales. Because of its algebraic nature, the standard deviation is the most useful for this purpose. The *standard scores* on tests $X_1, X_2, X_3 \cdots$ may then be defined as the deviations from the mean divided by the respective standard deviations, or $\frac{x_1}{\sigma_1}, \frac{x_2}{\sigma_2}, \frac{x_3}{\sigma_3} \cdots$. Like the

* Adapted from Miss Newcomer's figures in "Financial Statistics of Public Education in the United States, 1910–1920." The Macmillan Company, 1924.

coefficient of variation these scores are clearly abstract numbers, since they result from dividing a denominate number by a quantity of the same denomination.

It will be noted that the standard score of a pupil gives his relative position in the group in terms of a number of standard deviations above or below the mean. Thus, if the raw score be 120 with $M = 90$ and $\sigma = 10$, the standard score will be $\frac{120 - 90}{10} = +3$. If on another test this pupil scores 18 with $M = 12$ and $\sigma = 2$, his standard score will again be $+3$. His relative position in the distributions of both tests is clearly the same as shown by the standard scores.

Being abstract numbers, standard scores on several tests may be combined by addition. The only caution that needs to be observed is that the various distributions from which the original scores are taken for comparison shall be of the same general shape. For a very skewed distribution an observation one *S. D.* above the mean of the series is not comparable with a measurement one *S. D.* above the mean of a symmetrical group.

In order to illustrate the use of standard scores the following data resulting from seven different tests are presented:

TABLE 23. STANDARD SCORES OF A PUPIL ON SEVERAL TESTS

TEST	MEAN	*S. D.*	$X =$ SCORES OF A PUPIL	$x = X - M$*	$\frac{x}{\sigma}$
1	163	10.2	179	+16	+1.57
2	119	8.1	128	+9	+1.11
3	24	6.0	28	+4	+0.67
4	264	39.8	312	+48	+1.21
5	74	8.2	89	+15	+1.83
6	7.3	2.1	6	−1.3	−0.62
7	133	16.4	151	+18	+1.10
Total			893		6.87
Mean			127.6		0.98

*The deviations $x = X - M$ are first computed and then each is divided by σ as shown in the last column.

If a composite score of the seven tests is desired, it would not appear correct to add the scores on the separate tests, because they are in unlike units and undue weight would be given to extreme scores. The objection that unlike quantities should not be added is not a serious one because even horses, pigs, cows, and sheep may be added together to secure the total number of farmyard animals. This amounts, of course, to broadening the unit so as to include all items in the sub-classes of the total group. The objection against the extreme weighting of some scores may be more important, for a score of 6 on one test may represent a mental effort as serious as a score of 179 on another scale. Both of the above difficulties are overcome when standard scores are used, the only trouble being the amount of arithmetic involved and the presence of positive and negative scores.

For very careful work the best method for comparing measurements and forming composites is through the use of standard scores. Test scores are far from stable, however, and great precision in their treatment is not always desirable or necessary. In many composite tests the components may be added in the unweighted form with practically as good results as by the standard score or other methods of weighting. This is illustrated by the Terman Group Intelligence Test consisting of ten parts. The simple total of all points made was found to agree (correlate) almost perfectly with the composite formed by weighting each of the separate tests and adding them. There is considerable disagreement, of course, in the case of some scores, but when fifty to one hundred cases are taken these individual differences have little effect upon the net result, especially when the number of test items is fairly large and they are not extremely uneven in weighted value.

Aside from the question of precision it may be important to represent scores in the standard form for the purpose of clearer interpretation. By a very simple formula based on standard scores it is possible to transmute the results on any number of tests so that they all have the same mean and standard devia-

tion. This method, which is quite old,* is frequently rediscovered and appears from time to time in a slightly different form in psychological and educational journals.

Let X_1 and X_2 represent the scores on two tests expressed in any units, M_1 and M_2 the respective means, and σ_1 and σ_2 the corresponding standard deviations. We may now write

$$\frac{x_1}{\sigma_1} = \frac{x_2}{\sigma_2},$$

or

$$x_1 = \frac{\sigma_1}{\sigma_2} x_2.$$

Since $x = X - M$, this may also be expressed in the form

$$X_1 = M_1 + \frac{\sigma_1}{\sigma_2}(X_2 - M_2). \quad \left\{\begin{array}{c}\text{Transmutation formula}\\ \text{for comparable scores,}\\ \text{score form}\end{array}\right\} \qquad (23)$$

This is the desired transformation which, when applied to X_2, makes its mean and standard deviation equal to those of X_1. These properties are apparent from the preceding equation. Hence by applying this formula to each item in the series we may, without affecting the relative position of any value, change the series so that it will have any mean and standard deviation desired.

As an example we may select $M_1 = 50$ and $\sigma_1 = 10$, these being convenient numbers. By the application of the above transformation to any number of tests, they may be brought into direct comparison with the one selected as standard. Thus the series of X scores shown below may be transmuted into comparable T scores† by the relation

$$T = X_1 = 50 + \frac{10\sqrt{3}}{2}(X - 3),$$

or

$$T = 24.02 + 8.66\, X.$$

* Galton introduced comparable measures in the form of deviations from the median divided by the semi-inter-quartile range.

† These are similar to McCall's T-Scores. See William McCall, How to Measure in Education. The Macmillan Company, 1922. See also Chapter XII, section 8, of the present text.

X	f		T	f
5	10		67.32	10
4	20	$M_x = 3$	58.66	20
3	30	$\sigma_x = \frac{2}{\sqrt{3}} = 1.155$	50.00	30
2	20		41.34	20
1	10		32.68	10
	90			90

The distribution of T scores obviously has a mean of 50 and a standard deviation of 10. By similar transformations any number of series will have these same properties, so that the scores on all tests may be brought into direct comparison. Thus a score of 50 will always represent the performance of an individual at the mean, while 30 will represent the score of a person two standard deviations below the mean, etc. If such a scaling method were adopted it should, of course, be applied only to large groups of unselected children at different ages or grades. After the T scores have been worked out for the different tests, transmutation tables should be prepared so that any X score can be easily transformed into the corresponding T score.

8. The Measurement of Skewness

Whenever it becomes necessary to compare several distributions of varying degrees of asymmetry or skewness, some numerical measure of this property becomes desirable. Such a measure of skewness should be independent of the unit of measurement for the variable of the distribution. Thus for a distribution of heights, a representation of skewness is needed which will remain unchanged whether the measurements be made in inches or in centimeters.

One such measure may be obtained by the formula

$$S_k = \frac{(Q_3 - Md) - (Md - Q_1)}{Q} = \frac{Q_1 + Q_3 - 2\,Md}{Q}. \qquad \left\{\begin{array}{l}\text{Measure of}\\ \text{skewness}\\ \text{based on}\\ \text{quartiles}\end{array}\right\} \quad (24)$$

The skewness will thus be positive when the longer tail of the distribution is in the direction of the high values of the variable as shown in Fig. 28.

The lowest value given by (24) is clearly zero when the distribution is symmetrical. While a maximum value of 2 may be obtained with the formula, in actual practice results beyond the limits $\pm$ 1 are rare.

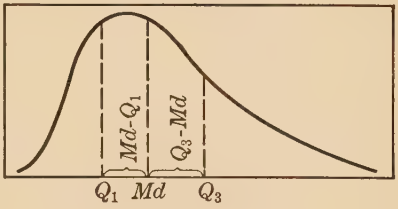


FIG. 28. A positively skewed distribution

A better measure of skewness is given by Pearson's formula,

$$S_k = \frac{M - M_o}{\sigma}, \left\{\begin{matrix}\text{Pearson's measures}\\ \text{of skewness}\end{matrix}\right\} \quad (25)$$

which also gives positive values for distributions of the type shown in Fig. 28. Owing to the fact that the true mode, M_o, is very difficult to determine, this formula may be replaced by another expression in which an approximate value for M_o is employed. Pearson has shown that for moderately skewed distributions, the relation between mode, mean, and median is given by

$$M_o = M - 3(M - Md).$$

Substituting this value for M_o in equation (25) we find

$$S_k = \frac{3(M - Md)}{\sigma}. \left\{\begin{matrix}\text{Approximate meas-}\\ \text{ure of skewness}\end{matrix}\right\} \quad (26)$$

As an example we may work out the degree of skewness in the distribution of I.Q.'s of Table 20, using formulas (24) and (26).

Using (24),

$$S_k = \frac{(100.77 - 89.31) - (89.31 - 77.437)}{11.67} = -.035.$$

Using (26),

$$S_k = \frac{3(89.28 - 89.31)}{16.86} = -.0053.$$

For this distribution the skewness, measured by either formula, is negative and slight.

EXERCISES

1. Calculate the mean deviation and the standard deviation for the following scores (ungrouped): 166, 159, 158, 151, 150, 142, 131, 126, 118, 101. ($M.D. = 17.0$; $S.D. = 19.7$. *Ans.*)

2. Compute the standard deviations for the frequency distributions of the data of Exercise 3, Chapter II.

($\sigma_o = 19.9$; $\sigma_c = 10.5$; $\sigma_t = 23.5$. *Ans.*)

3. Calculate the mean deviation, standard deviation, and quartile deviation for each of the problems of Exercise 1, Chapter VI.

	1	2	3	4	5	6	7	8	9
M.D.	6.85	2.79	1.40	11.43	16.63	4.61	5.25	1.33	1.43
S.D.	8.74	3.61	1.84	15.08	21.52	5.67	7.06	1.67	1.76
Q	5.94	2.125	1.125	7.875	11.29	3.85	4.31	1.03	1.30

Ans.

4. Calculate the coefficients of variation for the following distributions:

Monthly Salary in 1914	High-School Science Teachers	High-School English Teachers
$135–139.99	1	—
130–134.99	3	—
125–129.99	4	—
120–124.99	4	3
115–119.99	2	—
110–114.99	10	1
105–109.99	7	1
100–104.99	26	2
95–99.99	8	—
90–94.99	16	3
85–89.99	22	10
80–84.99	15	30
75–79.99	15	36
70–74.99	5	31
65–69.99	4	20
60–64.99	2	8
55–59.99	—	1
50–54.99	—	1
	144	147

Science: $M = \$94.83$, $\sigma = \$15.8$, $V = 16.7$
English: $M = \$77.67$, $\sigma = \$10.9$, $V = 14.0$

Ans.

The V's are more nearly alike than the σ's. Explain.

5. Verify the results in the following table:

COMPARISON OF FOREIGN-BORN GROUPS FOR DIFFERENT NUMBERS OF YEARS IN THE UNITED STATES IN TERMS OF THEORETICAL COMBINED SCALE OF INTELLIGENCE (ALPHA, BETA, AND ALL INDIVIDUAL EXAMINATIONS COMBINED)*

(Intervals are to be taken as 22–23 with class value 22.5 etc.)

COMBINED SCALE	YEARS IN UNITED STATES					TOTAL
	0–5	6–10	11–15	16–20	Over 20	
22	1.0	0.4	1.0	0.4	0.5	3.3
21	2.8	2.8	2.4	1.6	3.6	13.2
20	5.8	8.1	6.2	3.83	7.4	31.33
19	14.0	18.5	12.98	8.22	12.1	65.8
18	27.8	38.1	27.83	17.94	24.78	136.45
17	55.5	72.7	52.24	32.68	41.11	254.23
16	104.4	142.4	88.51	59.62	66.18	461.11
15	172.5	240.7	139.85	76.99	86.44	716.48
14	265.3	355.2	199.78	115.24	106.35	1,041.87
13	368.8	490.1	273.95	127.44	127.11	1,387.4
12	441.2	597.0	308.72	119.86	113.13	1,579.91
11	461.5	596.9	247.31	86.62	69.95	1,462.28
10	470.9	529.9	189.02	50.39	44.34	1,284.55
9	454.3	474.7	150.88	27.54	28.48	1,135.9
8	342.5	347.4	100.32	17.08	17.77	825.07
7	212.7	207.8	57.38	7.52	9.29	494.69
6	106.8	101.6	26.58	3.92	3.25	242.15
5	44.8	37.2	10.02	1.45	.86	94.33
4	16.4	14.5	3.74	.50	.25	35.39
3	4.7	4.3	1.03	—	—	10.03
2	1.5	1.3	.32	—	—	3.12
1	.4	.3	—	—	—	.7
Total	3,575.6	4,281.9	1,900.06	758.84	762.89	11,279.29
First quartile	9.36	9.75	10.66	11.94	12.15	9.98
Median	11.29	11.71	12.53	13.51	13.74	12.03
Third quartile	13.34	13.61	14.28	15.15	15.59	13.93
Quartile deviation	1.99	1.93	1.81	1.61	1.72	1.98

6. Work out the standard scores for the first five pupils on the three intelligence tests of Exercise 1, Chapter II, using the means and standard deviations already calculated.

Otis:	1.59	1.49	−.57	.09	−1.67
Chicago:	−.17	2.07	−.31	−.74	−1.36
Terman:	−.32	1.21	.28	−.83	−2.28 *Ans.*

* Data from Memoirs of the National Academy of Sciences, Vol. XV, p. 704.

7. Convert the following distribution into a series having mean = 85 and $S.D. = \sqrt{30}$.

Given Distribution		Transformed Distribution	
Class Value, X	f	T	f
80	1	96.62	1
70	2	92.75	2
60	3	88.88	3
50	8	85.00	8
40	3	81.13	3
30	2	77.26	2
20	1	73.39	1
	20		20

Transformation equation is $T = .3873\,X + 65.64$.

8. Derive formula (17) from (16).

CHAPTER VIII

THE PERCENTILE METHOD

1. Introductory

There is nothing essentially new in the method of percentiles, but the recent wide use of percentile scores, ranks, and curves in dealing with test data warrants a somewhat detailed account of this method.

It is hardly worth while to apply the percentile method in any form unless the data are sufficient in number to justify their representation in a frequency distribution. Percentiles are obtained in the same way as the median and quartile values which, as we have seen, are not well defined in the case of ungrouped items. Furthermore, the irregular nature of short series makes the percentile values unstable and of little practical significance. For these reasons the method will be discussed only in connection with frequency distributions.

2. Percentiles

A *percentile* is a value of the variable below which a given per cent of the frequencies lie. These values may be denoted by the symbol P_p, where the subscript p refers to the percentage of observations smaller than P_p. Thus P_{10}, P_{25}, P_{50}, and P_{82} are values such that 10, 25, 50, and 82 per cent of the cases lie below them.

From this definition it is apparent that the median is equal to P_{50} and that the quartile values Q_1 and Q_3 are equal respectively to P_{25} and P_{75}.

Formulas for the computation of percentile values may now be set up in a form similar to those used for the median.

Let p = the percentage of the cases smaller than P_p,
f_P = the frequency of the class where P_p occurs,
f_{up} and f_{do} = the frequency up to and down to the interval containing the required percentile,
$u.l.$ and $l.l.$ = the upper and lower limits of this interval, and
h and N = the size of the interval and sample as before.
The formulas then become

$$P_p = l.l. + \left[\frac{\frac{pN}{100} - f_{up}}{f_P}\right]h \qquad \left\{\begin{matrix}\text{Percentiles,}\\ \text{counting up}\end{matrix}\right\} \qquad (27a)$$

and

$$P_p = u.l. - \left[\frac{\left(\frac{100-p}{100}\right)N - f_{do}}{f_P}\right]h. \qquad \left\{\begin{matrix}\text{Counting}\\ \text{down}\end{matrix}\right\} \qquad (27b)$$

TABLE 24. ILLUSTRATING THE COMPUTATION OF PERCENTILES

AGE RECEIVED PH. D.	f	
45	3	
44	—	
43	3	
42	3	
41	1	
40	5	
39	9	
38	5	
37	5	
36	7	$308 = f_{do}$
35	7	
34	10	
33	13	
32	17	
31	29	
30	42	
29	31	
28	27	
27	37	
26	54	
25	38	$= f_P$
24	29	
23	14	
22	7	$54 = f_{up}$
21	2	
20	2	
	400	

$\frac{pN}{100} = .20 \times 400 = 80$

$P_{20} = 24.5 + \frac{80 - 54}{38} \times 1$

$= 24.5 + .684$

$= 25.184$

Check:

$P_{20} = 25.5 - \frac{320 - 308}{38} \times 1$

$= 25.5 - .316$

$= 25.184$

In order to illustrate the use of formulas (27 a) and (27 b), the complete calculation for P_{20} is given in Table 24. It should be noted that the ages are given at class values, the intervals being 44.5–45.5, 43.5–44.5, etc. The check should be used until the student is confident of the accuracy of his calculations.

By similar computations, the values for P_{10}, P_{20}, up to P_{90} may be obtained and set down as follows:

$$\begin{array}{lll} P_{10} = 24.02, & P_{40} = 26.88, & P_{70} = 30.43 \\ P_{20} = 25.18, & P_{50} = 28.13, & P_{80} = 31.97 \\ P_{30} = 26.02, & P_{60} = 29.47, & P_{90} = 35.64 \end{array}$$

These percentiles divide the series into ten equal parts so that a given age may be readily located in any part of the distribution. Thus if a man received his Ph.D. at twenty-six, it is at once apparent that 30 per cent of the men were younger than he when they took this degree. Similarly, a man who received the degree at thirty-two was among the oldest fifth of the entire group.

The above method for obtaining percentile values is the most direct and accurate one. The same results may be obtained more easily, however, by making use of the cumulative frequency curve as described in Chapter II. The computation in this case is graphical and the accuracy of the results will depend upon the construction and use of the drawing. When adding in from the lower end of the series, the cumulative frequency distribution for ages may be arranged as shown in Table 25 on page 130.

The plot of these data is shown in the cumulative frequency curve of Fig. 29 on page 131. The p scale on the right is made by dividing the total cumulative frequency scale into 100 equal parts. In order to obtain any percentile value graphically it is only necessary to find the required percentile index p, on the p scale, move to the left from this point until the curve is reached, and then drop down vertically to the percentile value on the horizontal scale.

TABLE 25. CUMULATIVE FREQUENCY DISTRIBUTION FOR DATA OF TABLE 24

AGE	FREQUENCY LESS THAN GIVEN AGE
45.5	400
44.5	397
43.5	397
42.5	394
41.5	391
40.5	390
39.5	385
38.5	376
37.5	371
36.5	366
35.5	359
34.5	352
33.5	342
32.5	329
31.5	312
30.5	283
29.5	241
28.5	210
27.5	183
26.5	146
25.5	92
24.5	54
23.5	25
22.5	11
21.5	4
20.5	2

Fig. 29 has been drawn in the form of a polygon, consisting of straight lines between the cumulative frequency points. While it is sometimes legitimate to smooth in the points by a free-hand or fitted curve, the student had better confine himself to the use of the polygon until he has made a special study of the subject of smoothing.

Although greater precision may be obtained by the use of the direct method of computing percentile values, the equivalence of the two procedures may be readily seen. Because of the manner in which the cumulative frequency curve is constructed the value of the ordinate gives the total frequency below the corresponding abscissa. Thus 54 frequencies lie below 24.5, and 92 frequencies lie below 25.5. By joining these points with

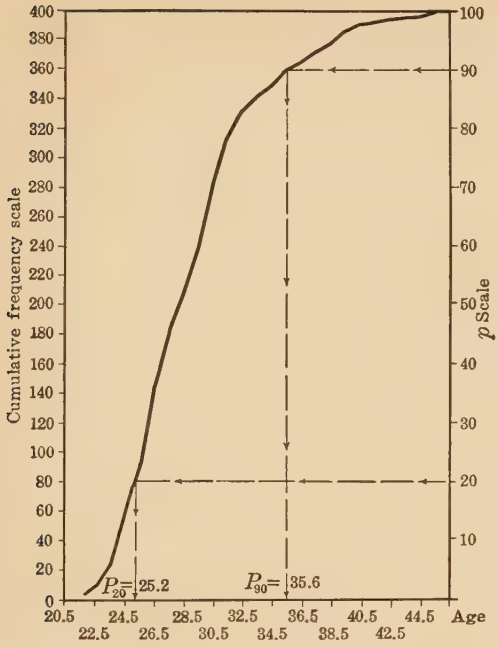


FIG. 29. Cumulative frequency curve for ages at which Ph.D.'s were received

a straight line as in Fig. 30, it is assumed that the increase in cumulative frequencies AB is directly proportional to the part of the interval up to the corresponding point P on the horizontal scale. In this case the frequency at P is 20 per cent of the observations, or 80, so that $AB = 26$. The value of x is therefore found from the proportion

$$\frac{x}{1} = \frac{26}{38} = .68.$$

Adding this result to 24.5, the lower limit of the interval, gives 25.18, or exactly the same result as by the direct method of calculation.

3. PERCENTILE CURVES

A percentile curve may be made by plotting the series of values such as those worked out in section 2. The ordinate is the value of the percentile, while the abscissa is the percentile index p. Such curves should be distinguished from cumulative frequency curves where integrated frequency is

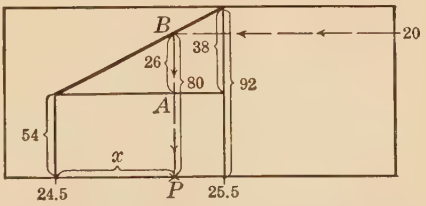


FIG. 30. Enlargement of a portion of the cumulative frequency curve to illustrate the calculation of P_{20}

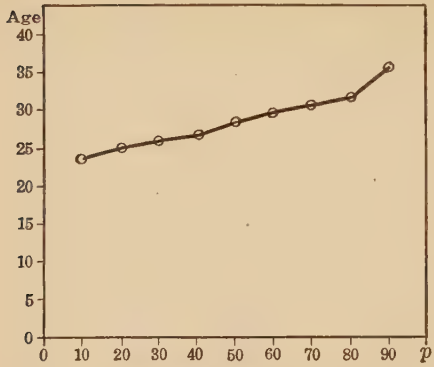


FIG. 31. Percentile curve for Ph.D. data

represented by the ordinate and values of the variable by the abscissa. Both of these are often called percentile curves, but it is more in harmony with mathematical convention to name a curve according to what is represented by the ordinate, and this is the basis for the above distinction.

Inspection of Figs. 29 and 31 shows that the two curves are essentially different in form, one being reversed in curvature from the other. The percentile curve has been termed by Francis Galton an *ogive*.

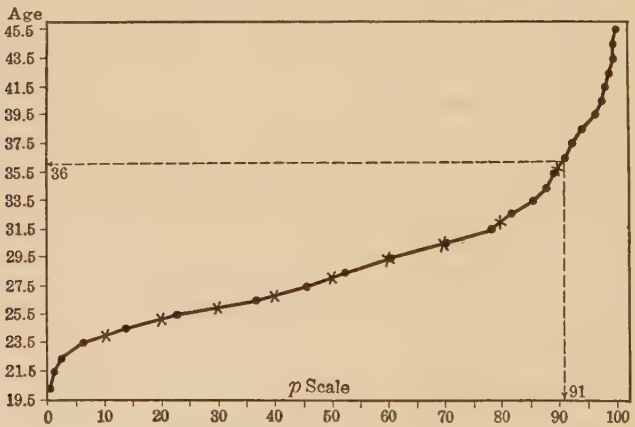


FIG. 32. Percentile or ogive curve for Ph.D. data

Another method for constructing such ogives is by means of the cumulative frequency distribution.* The difference between this method and the one just shown is that values of the ends

* This method is used by Otis in connection with the tests published by the World Book Company.

of the class intervals are now plotted over computed p-scale values instead of finding the percentiles at given p values, 10, 20, 30, etc., and plotting them.

The calculation is very simple, consisting of a series of cumulative frequency percentages for the values of p at the ends of the intervals. In Table 26 these results are set forth for the Ph. D. data. The computation may be done most readily by setting the reciprocal of 400 in the calculating machine and multiplying it into the series of cumulative frequencies. The curve as shown in Fig. 32 is now constructed by plotting the values from Table 26 and connecting the resulting points. These points are indicated by dots, while those previously calculated

TABLE 26. ILLUSTRATING CUMULATIVE FREQUENCY PERCENTAGES

AGE	fc = FREQUENCY LESS THAN GIVEN AGE	p = PERCENTAGE FREQUENCY LESS THAN GIVEN AGE $= \frac{fc}{N} \times 100$
45.5	400	100.0
44.5	397	99.3
43.5	397	99.3
42.5	394	98.5
41.5	391	97.8
40.5	390	97.5
39.5	385	96.3
38.5	376	94.0
37.5	371	92.8
36.5	366	91.5
35.5	359	89.8
34.5	352	88.0
33.5	342	85.5
32.5	329	82.3
31.5	312	78.0
30.5	283	70.8
29.5	241	60.3
28.5	210	52.5
27.5	183	45.8
26.5	146	36.5
25.5	92	23.0
24.5	54	13.5
23.5	25	6.3
22.5	11	2.8
21.5	4	1.0
20.5	2	0.5

from the simple distribution are given for comparison by small crosses. It is obvious that these latter values could have been obtained graphically by making use of the ogive formed by the dots.

4. Use of Percentile Curves

A review of the foregoing paragraphs will show that percentile values may be calculated in three ways. They may be computed numerically from formulas (27 a) and (27 b), making use of the simple frequency distribution; they may be calculated graphically from the cumulative frequency curve; and finally they may be obtained graphically from the ogive, or percentile curve.

The particular method to be used depends upon the adequacy of the data, the number of percentiles required, and the accuracy needed. Unless the data are fairly plentiful (one hundred or more cases), the graphical methods are usually not as expeditious as the use of the formulas. Furthermore, if only the median and quartiles are required, it does not pay to throw the data into cumulative or ogive form in order to obtain them. Finally, if considerable accuracy be needed in the result, the numerical method is far superior to the others.

In case a number of percentiles are required and the data are sufficient in number, either of the above graphical methods may be employed to advantage, where only fairly accurate results are needed. If the total number of cases gives a convenient quotient when divided by 100, the p scale of the cumulative curve may be readily constructed, and this method is probably the better to use. For most problems, however, the total frequency is an awkward number such as 371, so that the graphical construction of the p scale becomes difficult. It is therefore usually best in constructing both the cumulative frequency curve and the ogive, to use the percentage frequencies as shown in Table 26.

Another use of percentile curves is in the comparison of two series. As an example, one of Otis's graphs is shown in Fig. 33. The two curves shown have been smoothed free-hand, but as

already indicated the subject of smoothing is beyond a course such as this, and the student will ordinarily do better to take the data at their face value in drawing such percentile curves.

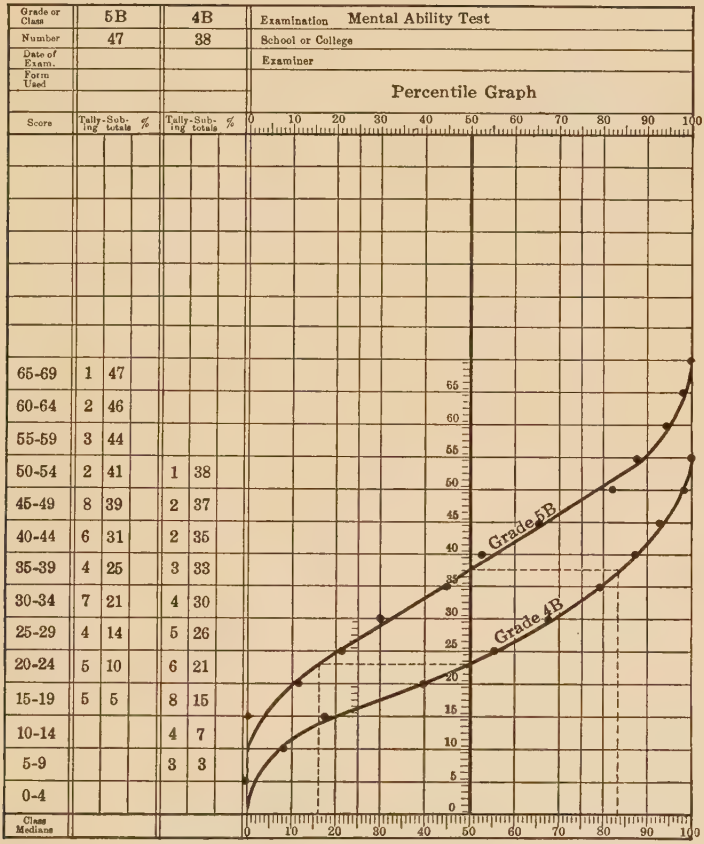


FIG. 33. Illustrating the method of drawing a percentile curve*

Otis† points out that the scores in Grade 5B are appreciably higher than those of Grade 4B, but that on the whole the dis-

* From Arthur S. Otis, Statistical Method in Educational Measurement. World Book Company, Yonkers-on-Hudson, New York, 1925.

† A. S. Otis, Statistical Method in Educational Measurement, p. 87. World Book Company, 1925.

tributions of scores of the two grades overlap very markedly. He goes on to show that "a convenient way to express the overlapping is to state the per cent of scores of Grade 4B that exceed the median score of Grade 5B, or to state the per cent of scores in Grade 5B that fall below the median score of Grade 4B. Thus, by finding the point on the 4B curve having a height representing a score of 37 (the median score of Grade 5B), we find that the upper 17 per cent of the scores on Grade 4B, as indicated by the curve, are above the median score of Grade 5B. The dotted lines indicate the solution." Otis also shows that such curves are convenient for finding and comparing various percentile values. Thus the pupils at the 10 and 90 percentiles in the two groups differ less widely than do the corresponding median pupils. This is shown by the vertical distance between the curves. There is, however, a certain amount of optical illusion in such comparisons which makes the curves appear to bulge apart in the middle.

5. Percentile Ranks

The percentile curve is also useful in determining graphically what are known as percentile ranks. These are the p values on the horizontal scale for such a curve. *The percentile rank of a given score is therefore the per cent of the observations below that score in the distribution.* In obtaining such ranks from the ogive it is only necessary to find the given score on the vertical scale, run across horizontally until the curve is met, and then drop down at right angles to the required p value, or percentile rank. As an example, making use of Fig. 32, let it be required to find the rank of a man who received his Ph. D. at 36. The result, as shown, is a percentile rank of about 91. This means that out of one hundred such men nine were older when they received this degree.

It may be noted that for percentile ranks 100 is high and 1 is low, which is contrary to the ordinary practice of assigning 1 to

the highest score in a series. Since the p values should naturally increase with an increasing value of the variable, this reversal seems justifiable in the case of percentile ranks, which need not be confused with ordinary ranks if properly specified.

A formula for percentile ranks may be derived, making use of numerical rather than graphical interpolation as illustrated above. This method will be first illustrated by an enlargement of the portion of the ogive including age 36 as shown in Fig. 34. From similar triangles it is apparent that $\frac{a}{36-35.5} = \frac{91.5-89.75}{1}$, or $a = .875$. The required percentile rank is therefore $89.75 + .875 = 90.625$. Such great accuracy as this is, of course, rarely necessary, and the final result may here be written 90.6, or possibly 91, as before.

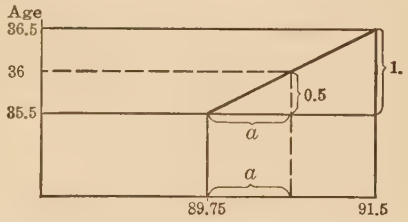


FIG. 34. Illustrating linear interpolation with the ogive

The formula for percentile ranks may now be set up by letting

X = the value of the given score,
R_x = its percentile rank,
$l.l.$ = lower limit of the interval containing X,
R_u and R_l = the percentile ranks of the upper and lower limits of this interval, and
h = width of the class interval.

We therefore have

$$R_x = R_l + \frac{R_u - R_l}{h}(X - l.l.). \quad \left\{\begin{array}{c}\text{Percentile}\\ \text{rank formula,}\\ \text{form 1}\end{array}\right\} \quad (28)$$

As an illustration of the use of this formula we may compute once more the percentile rank for the age 36. From Table 26, $R_u = \frac{100 \times 366}{400} = 91.5$, and $R_l = \frac{100 \times 359}{400} = 89.75$. We therefore obtain

$$R_{36} = 89.75 + \frac{91.5 - 89.75}{1}(36 - 35.5) = 90.625 = 91.$$

A more direct and usually more convenient formula for percentile ranks may be obtained by letting

f_x = frequency of the interval where X occurs,

and f_{up} = frequency up to this interval.

We may then write

$$R_x = \frac{100[f_x(X - l.l.) + (f_{up})h]}{Nh} \quad \left\{\begin{matrix}\text{Percentile rank}\\ \text{formula, form 2}\end{matrix}\right\} \quad (29)$$

Thus from Table 24 the frequency at age 36 is $f_x = 7$, while the frequency up to this interval is 359. Substituting these values in formula (29), we find that

$$R_{36} = \frac{100[7(36 - 35.5) + 359 \times 1]}{400 \times 1} = 90.625,$$

as before. The student should show that formulas (28) and (29) are equivalent.

Instead of finding the percentile rank of X it is often sufficient to find the percentile rank of the class value of the interval where X occurs. In this case, formula (29) reduces easily to

$${}_cR_x = \frac{50f_x}{N} + \frac{100(f_{up})}{N} = \frac{50f_x}{N} + R_l. \quad \left\{\begin{matrix}\text{Class value}\\ \text{rank}\end{matrix}\right\} \quad (30)$$

Applying this form to age 36, we again find ${}_cR_{36} = 90.625$, since 36 happens to be the class value of the interval. The rank 91 would be used for any age between 35.5 and 36.5, according to this last approximation, and this is often sufficiently accurate with a narrow class interval.

The percentile ranks of a set of scores often make a very convenient record for administrative use. This may be illustrated in the case of a group of graduate students who were given an intelligence test. The gross scores were not used because of the lack of suitable norms for such groups. By converting the scores of the tests into percentile ranks, the relative position of each student in the group could be seen at a glance. Thus John Doe's graduate record might appear as follows:

General average of college marks	A−
Estimated fitness for research	Excellent
Personality	Pleasing
Experience in college teaching	3 years
Age	25
Percentile rank in intelligence test	97

The youth, high scholarship, and personality of this man are indications of future success in college teaching. His percentile rank of 97 is additional evidence in this respect, because in a group of 119 students, he was exceeded by only 3 per cent in general mental alertness.

Inasmuch as a very accurate rank for such purposes is not required, the graphical method of determination from the percentile curve may be conveniently used. An error of 1 per cent in the percentile rank of a student will make no difference in the administrative interpretation of the test result, and this degree of accuracy may be easily obtained from the ordinary free-hand graph.

While percentile ranks will furnish the medians and quartiles of the original distribution of scores, the ranks should not be treated like actual scores. In combining scores from several tests, for example, it would not be legitimate to add the raw scores from certain tests to other scores expressed in the form of percentile ranks. The distribution of such ranks will approximate a long rectangle, the standard deviation of which is of doubtful significance. It is therefore much better to keep the data in their original form for most purposes and to convert the items into percentile ranks only for such uses as those which have been described above.

In general the whole percentile method is cruder but sometimes more convenient than methods in which the raw scores are employed directly. Percentile curves and ranks are in extensive use at present, but for careful analytical work it is usually best to employ methods based on the actual rather than on the relative values of the scores.

EXERCISES

1. Calculate the nine deciles by formula (27) for the distribution of science teachers' salaries given in Exercise 4 of Chapter VII. ($P_{10} = \$76.13$; $P_{20} = \$80.93$; $P_{30} = \$85.50$; $P_{40} = \$88.77$; $P_{50} = \$92.81$; $P_{60} = \$99.63$; $P_{70} = \$102.65$; $P_{80} = \$106.57$; $P_{90} = \$114.80$. *Ans.*)

2. Construct a cumulative frequency curve for the data of Exercise 1, and check the above deciles by graphical computation, checking by the graphical method.

3. Work out the nine deciles for the distribution of the fourth-year high-school group from the table in Exercise 3, Chapter VI. ($P_{10} = 66.74$; $P_{20} = 81.39$; $P_{30} = 92.76$; $P_{40} = 102.23$; $P_{50} = 110.91$; $P_{60} = 119.01$; $P_{70} = 128.84$; $P_{80} = 138.98$; $P_{90} = 152.12$. *Ans.*)

4. Construct a percentile curve from the values obtained in Exercise 3.

5. Calculate a table of class-value percentile ranks, using formula (30) and the data of Exercise 3. Check by the cumulative frequency curve.

6. Compute by formula (29) the percentile ranks for the following scores: 167, 35, 171, 81, and 104, using the distribution of Exercise 3. (96.4, 1.8, 97.5, 19.7, 42.0. *Ans.*)

7. Prove that formulas (28) and (29) are equivalent.

CHAPTER IX

LINEAR CORRELATION WITH QUANTITATIVE SERIES

1. The Meaning of Correlation

Correlation is sometimes defined as the concomitant variation of two traits. This definition may be illustrated by the scores of fifty pupils on the Otis and Chicago group intelligence tests listed in Exercise 1, Chapter II. In running through the pairs of scores for each pupil, it will be noted that a high score on one test is usually associated with a high score on the other, while a low score on one test tends to be paired with a correspondingly low score on the second test.

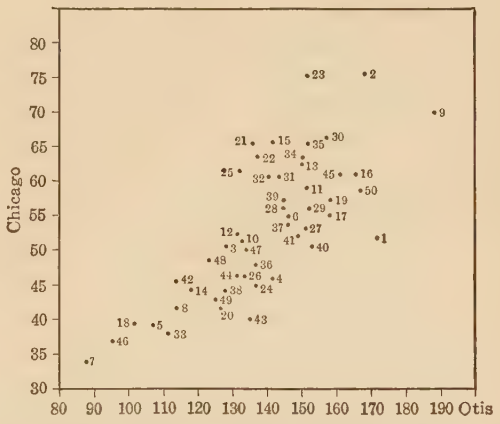


Fig. 35. Scatter diagram of the Otis and Chicago test scores. (The numbers identify the scores given in Exercise 1, Chapter II)

This relationship, or correlation, is brought out more clearly by means of a *scatter diagram,* which is merely a plot of the associated pairs of scores as shown in Fig. 35.

There is a general tendency in this diagram for the points to form a straight band across the graph, and this furnishes graphical evidence of *linear correlation.* The narrower the band and the more closely the points cluster along a straight line, the higher such correlation becomes.

In Fig. 36 another scatter diagram is shown, but the band in this example forms a distinct curve. The correlation in this case is regarded as non-linear, but like linear correlation the relationship between the two variables becomes closer as the points form a narrower and narrower band, finally approximating a single-valued mathematical function.

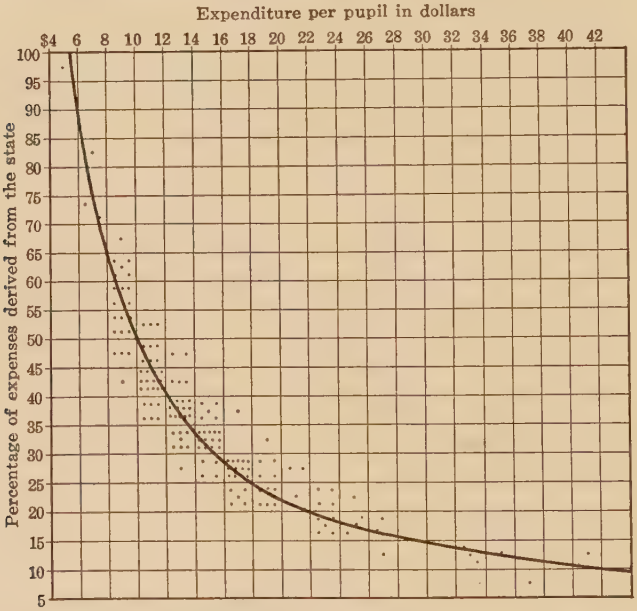


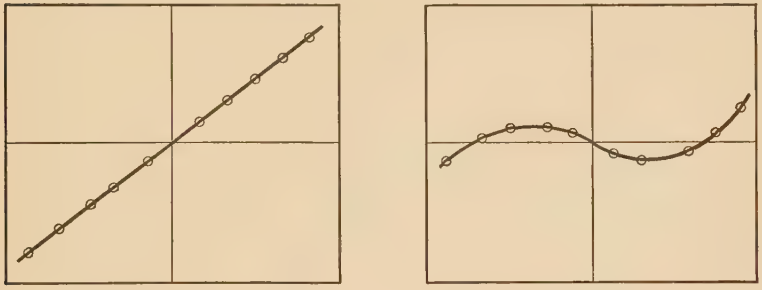
FIG. 36. Showing the relationship between per-pupil expenditure and percentage of total school expenditure derived from the state. (Data supplied by Dr. R. E. Wager)

Perfect correlation is reached when all of the points in the scatter diagram fall exactly on a curve.* Two examples of such relationship are shown in Fig. 37, one for linear and one for non-linear correlation. With observed data, perfect correlation is, of course, impossible but very close approximations are often reached in verifying physical laws such as, stress $= k \times$ strain.

* It will be remembered that *curve* is a general expression for the designation of both linear (straight line) and non-linear functions.

In view of the above discussion we may now define *correlation as the tendency for two observed variables to be related in the form of a single-valued mathematical function*, or, more briefly, as the tendency toward single-valued functionality. A single-valued function is one such that for any value of the argument only one value of the function results.

The present chapter will deal with linear correlation for quantitative series, while Chapter X will be devoted to the measurement of curvilinear relationship. For both types of correlation



Perfect linear correlation Perfect non-linear correlation

FIG. 37. Illustrating two types of perfect correlation

it is possible to express the degree of association in numerical terms, and obtain an equation for the mathematical curve which most closely approximates the data.

2. THE PRODUCT-MOMENT CORRELATION COEFFICIENT

In Fig. 38 on page 144 it will be noted that the plane has been divided into four quadrants by erecting perpendiculars at the means on the two scales. Designating these quadrants in the usual way, it appears that points located in the first and third quadrants will tend to produce high correlation, while points located in the other two quadrants will tend to reduce the amount of such correlation. When the points are scattered randomly over the plane, the correlation will approach zero.

A measure of relationship might then be obtained by considering the products of the deviations from the means, that is, $x = X - M_x$, and $y = Y - M_y$. For a point such as P_1, the product xy will be positive, while for P_2 in the second quadrant, the xy product will be negative, etc. The average of all such pair-products might be used to measure correlation, were it not for the fact that the deviations are expressed in the units of the respective scales. In order to overcome this difficulty it is only necessary to use standard scores, that is, $\frac{x}{\sigma_x}$ and $\frac{y}{\sigma_y}$, and the resulting product average will then become a pure number. Thus for N pairs of associated points the *product-moment correlation coefficient* becomes

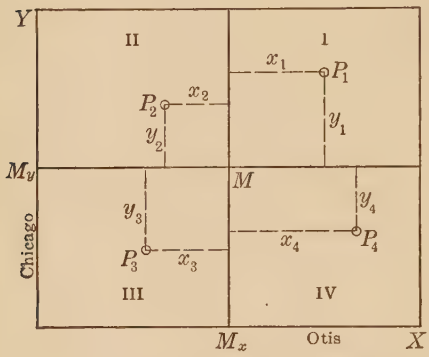


FIG. 38. Illustrating product moments in four quadrants

$$\left(\frac{x_1}{\sigma_x}\frac{y_1}{\sigma_y} + \frac{x_2}{\sigma_x}\frac{y_2}{\sigma_y} + \frac{x_3}{\sigma_x}\frac{y_3}{\sigma_y} + \cdots + \frac{x_N}{\sigma_x}\frac{y_N}{\sigma_y}\right) \Big/ N.$$

Representing this coefficient by the symbol r, and denoting the sum in the usual way, we have

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}, \quad \left\{\begin{array}{c}\text{Product-moment}\\ \text{correlation coefficient,}\\ \text{original form}\end{array}\right\} \quad (31)$$

or

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2\ \Sigma y^2}}, \quad (32)$$

since

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} \text{ and } \sigma_y = \sqrt{\frac{\Sigma y^2}{N}}.$$

The product-moment coefficient is thus the arithmetical mean of the pair-products of associated standard scores. The above

formula was developed originally by Karl Pearson,* who based his proof upon the product-moment function of Bravais, and a method of Galton's closely related to the above standard scores. It is therefore often called the Pearson correlation coefficient.

If all the points in the scatter diagram lie on a straight line, the equation of this line through the origin (Fig. 39) will

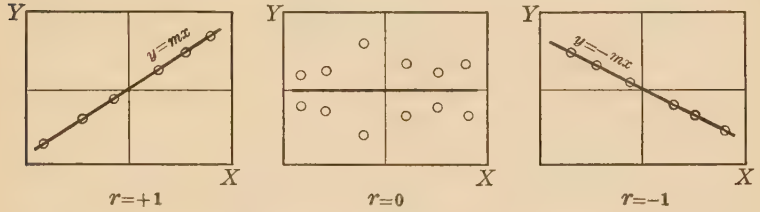


FIG. 39. Illustrating extreme variations in the correlation coefficient

be $y = \pm mx$, where $\pm m$ is the slope of the line. The value for the correlation coefficient in such a case may now be determined by noting that $\sigma_y = m\sigma_x$,

since $$\sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{\Sigma m^2 x^2}{N}},$$

and that $$\Sigma xy = \pm m\Sigma x^2 = \pm mN\sigma_x{}^2.$$

We may then write

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \pm\frac{mN\sigma_x{}^2}{mN\sigma_x{}^2} = \pm 1.$$

In case all the points lie on a horizontal line with a zero slope, the value for r becomes indeterminate, that is, $\frac{0}{0}$. A symmetrical arrangement of the points about such a line, however, will give zero correlation, as shown in Fig. 39, because the quantity Σxy is zero while N, σ_x, and σ_y are not zero. With actual data, therefore, the correlation coefficient may range in value from -1 to 1.

* A full discussion of the history of correlation is given by Pearson in *Biometrika*, Vol. XIII (1920). Here Pearson assigns most of the credit to Galton and minimizes the significance of his own important contribution.

3. Computation of the Correlation Coefficient with Ungrouped Items

While formulas (31) and (32) are useful for calculating the correlation coefficient, the arithmetic may become rather tedious on account of the fact that the deviations x and y are usually given in the form of decimals which would need to be multiplied and squared. In order to overcome this difficulty an alternative formula will be given.

Remembering that $x = X - M_x$, and that $y = Y - M_y$, we may write

$$xy = XY - YM_x - XM_y + M_xM_y,$$

and
$$\Sigma xy = \Sigma XY - M_x\Sigma Y - M_y\Sigma X + M_xM_y\Sigma(1)$$
$$= \Sigma XY - NM_xM_y,$$

since $\Sigma X = NM_x$, $\Sigma Y = NM_y$, and $\Sigma(1) = N$.

From the chapter on dispersion it is also evident that

$$\sigma_x = \sqrt{\frac{\Sigma X^2}{N} - M_x^2},$$

and that
$$\sigma_y = \sqrt{\frac{\Sigma Y^2}{N} - M_y^2}.$$

Substituting these values in equation (31) gives the desired formula,

$$r = \frac{\Sigma XY - NM_xM_y}{\sqrt{(\Sigma X^2 - NM_x^2)(\Sigma Y^2 - NM_y^2)}}. \quad \left\{\begin{array}{l}\text{Correlation coefficient}\\ \text{(based on raw scores)}\end{array}\right\} \quad (33)$$

This expression, although more complicated in form than formula (32), is generally preferable to the latter because it involves the integral scores X and Y rather than the deviations x and y as decimals. By designating the total as T, it is evident that $T = NM$, and the above formula may also be written

$$r = \frac{\Sigma XY - T_xM_y}{\sqrt{(\Sigma X^2 - T_xM_x)(\Sigma Y^2 - T_yM_y)}}. \quad \left\{\begin{array}{l}\text{Correlation coefficient}\\ \text{equivalent to (33)}\end{array}\right\} \quad (34)$$

It should also be noted that when the variables are measured from arbitrary origins, A_x and A_y, the last two formulas may be

applied to $X' = X - A_x$, and to $Y' = Y - A_y$. This follows at once from the fact that

$$M'_x = M_x - A_x, \text{ and } M'_y = M_y - A_y,$$

so that $x' = x$ and $y' = y$.

The primed variables may then replace the unprimed variables throughout in formulas (33) and (34), which means that the correlation remains unchanged when any numbers A and B are subtracted from X and Y, respectively.

The above formulas will now be applied to a short problem with the data in the form of listed pairs of associated scores on two tests, X and Y. The material is too scanty to have any practical value, but has been chosen because the arithmetic is short and the attention may be fixed on the form of computation.

TABLE 27. ILLUSTRATING THE COMPUTATION OF THE PRODUCT-MOMENT CORRELATION COEFFICIENT BY DEVIATIONS FROM THE MEANS

PUPIL	SCORE ON X	SCORE ON Y	x	y	x^2	y^2	xy
A	76	17	− 1.9	− 0.2	3.61	0.04	+ 0.38
B	74	15	− 3.9	− 2.2	15.21	4.84	+ 8.58
C	82	14	+ 4.1	− 3.2	16.81	10.24	− 13.12
D	63	12	− 14.9	− 5.2	222.01	27.04	+ 77.48
E	74	18	− 3.9	+ 0.8	15.21	0.64	− 3.12
F	91	19	+ 13.1	+ 1.8	171.61	3.24	+ 23.58
G	86	20	+ 8.1	+ 2.8	65.61	7.84	+ 22.68
H	82	23	+ 4.1	+ 5.8	16.81	33.64	+ 23.78
I	79	20	+ 1.1	+ 2.8	1.21	7.84	+ 3.08
J	72	14	− 5.9	− 3.2	34.81	10.24	+ 18.88
	77.9 (Mx)	17.2 (My)			562.90 (Σx^2)	105.60 (Σy^2)	162.20 (Σxy)

In applying formula (32) it is first necessary to obtain the means and deviations from the means for the two series, the latter being given in the columns x and y above. The squared and product terms are then formed and the sums Σx^2, Σy^2, and Σxy obtained. The value for the coefficient then becomes

$$r = \frac{162.2}{\sqrt{562.9 \times 105.6}} = \frac{162.2}{\sqrt{59442.24}} = \frac{162.2}{243.8} = .665.$$

When using formula (33) or (34), it will be found convenient to reduce the scores before calculating the necessary quantities. Subtracting 70 from each of the X scores and 15 from each of the Y scores gives the X' and the Y' series shown in the following illustration (data from Table 27).

TABLE 28. ILLUSTRATING THE COMPUTATION OF THE PRODUCT-MOMENT CORRELATION COEFFICIENT BY DEVIATIONS FROM ASSUMED MEANS

PUPIL	X'	Y'	$(X')^2$	$(Y')^2$	$(X'Y')$
A	6	2	36	4	12
B	4	0	16	0	0
C	12	−1	144	1	−12
D	−7	−3	49	9	21
E	4	3	16	9	12
F	21	4	441	16	84
G	16	5	256	25	80
H	12	8	144	64	96
I	9	5	81	25	45
J	2	−1	4	1	−2
Totals	79	22	1187	154	336
M'	7.9	2.2	$624.1 = T'xM'x$	$48.4 = T'yM'y$	$173.8 = T'xM'y$
$T'M'$	624.1	48.4	562.9	105.6	162.2

$T'xM'y = T'yM'x = 173.8.$

The squaring and multiplying may now all be done mentally and the computation arranged as in the foregoing scheme. Using formula (34) we then have

$$r = \frac{162.2}{\sqrt{562.9 \times 105.6}} = +.665,$$

as before.

Of the two types of calculation shown above, the second is usually the easier, although both become rather tedious with a long series. Formula (33) and occasionally formula (32) are recommended for short series of, say, 20 to 30 pairs of scores, which do not warrant the use of a frequency table.

As a warning to the student, it may be noted that correlations based upon such a small number of cases are not of much significance. In experimental work, however, problems of this sort do arise, and it would then be convenient to employ the above methods.

4. The Computation of the Correlation Coefficient for a Frequency Table

A two-way frequency table, or correlation table, may be made by noting the frequencies which occur in the *cells* bounded by certain class limits on the two scales. Thus, by taking class intervals of 79.5–89.5, 89.5–99.5, 99.5–109.5, etc. for the Otis test, and 29.75–34.75, 34.75–39.75, etc. for the Chicago test, a scheme for tabulation may be set up as shown in Table 29 (see data from Exercise 1, Chapter II). Instead of recording a pair of scores as a point on a scatter diagram, it is only necessary to make a tally in the cell within which this pair of scores must lie. All frequencies in a given cell are then assumed to have the class values of that cell. For example, the scores of the first pupil are 171 and 52. This pair of scores is recorded by a tally

Table 29. Correlation Table for the Otis and Chicago Test Scores*

Chicago Score	Otis Score 80	90	100	110	120	130	140	150	160	170	180	190	Total
80								/	/				2
75												/	1
70						/	/	//					4
65						//	//	//	//				8
60							///	////	/				8
55					/	///	//	//		/			9
50				/	/	////	/			First pair of scores			7
45				//	///	/							6
40		/	//	/									4
35	/												1
30													
Total	1	1	2	4	5	11	9	11	4	1	1		50

* The exact class limits have not been set down in the table, but these should be kept in mind in tabulating the frequencies and in subsequent calculations for the means.

in the cell indicated by 169.5–179.5 on Otis and 49.75–54.75 on Chicago, with the resulting class values 174.5 and 52.25. It will be noted that similar errors (differences between class values and observed values) appear throughout the table, but that the combined effect of these upon the correlation coefficient will not be large if the class intervals are fairly numerous on both scales, for example, from 10 to 20 intervals as in the case of the simple frequency distribution.

Unless a mechanical sorting device is available the best way to make a correlation table is with the aid of the small data tickets described in Chapter II. These may be sorted into a simple frequency distribution according to one of the variables, and each of the sub-piles then sorted for a distribution of the associated variable. It will be found convenient to write down the class limits on small slips of paper, laying these out in a row on a long table. The cards are then sorted into piles and the work verified by running through each one. These piles may then be secured with rubber bands and a new series of class intervals prepared for the next variable. As soon as each pile has been sorted in this way, the results may be tabulated in the appropriate column on a sheet of square-ruled paper or on a special form to be described below.

In calculating the correlation coefficient for a frequency table it will be necessary to modify formula (33) so as to bring in the frequency notation. Let

f_x = the frequency of a column of type x,

f_y = the frequency of a row of type y,

f_{xy} = the frequency of a cell common to such a column and row,

d_x and d_y = the deviations in class intervals from the assumed means on the two scales,

h and k = the widths of the class intervals for the variables X and Y, respectively, and

X' and Y' = the variables measured from arbitrary origins, A_x and A_y.

It is now evident that $d_x = \frac{X'}{h}$ and $d_y = \frac{Y'}{k}$,

so that $\Sigma X'Y' = \Sigma d_x d_y hk = (\Sigma f_{xy} \cdot d_x d_y)hk$,

f_{xy} being merely a symbol of operation. In a similar way it appears that

$$NM'_x M'_y = \frac{(\Sigma f_x d_x)(\Sigma f_y d_y)}{N} hk, \quad \Sigma(X')^2 = (\Sigma f_x d_x^2)h^2,$$

and $$N(M'_x)^2 = \frac{(\Sigma f_x d_x)^2 h^2}{N}.$$

Substituting these values in formula (33) gives

$$r = \frac{\Sigma f_{xy} d_x d_y - \frac{(\Sigma f_x d_x)(\Sigma f_y d_y)}{N}}{\sqrt{\left[\Sigma f_x d_x^2 - \frac{(\Sigma f_x d_x)^2}{N}\right]\left[\Sigma f_y d_y^2 - \frac{(\Sigma f_y d_y)^2}{N}\right]}} = \frac{a}{\sqrt{bc}}, \quad (35)$$

{Correlation coefficient for distribution table}

from which it follows that the correlation coefficient is quite independent of the magnitudes of the class intervals and of the units of measurement. The three principal terms in this expression have been designated as a, b, and c for convenience.

The complete calculation with formula (35) is illustrated in Table 30 on page 152 for the Otis-Chicago data.

$$a = 170 - \frac{15 \times 24}{50} = 170 - 7.2 = \boxed{162.8}$$

$$b = 210 - \frac{(24)^2}{50} = 210 - 11.52 = \boxed{198.48}$$

$$c = 225 - \frac{(15)^2}{50} = 225 - 4.5 = \boxed{220.5}$$

$$r = \frac{a}{\sqrt{bc}} = \frac{162.8}{\sqrt{198.48 \times 220.5}} = \frac{162.8}{\sqrt{43764.84}} = \frac{162.8}{209.2} = .778.$$

By four-place logarithms,

$$\log b = 2.2978$$
$$\log c = 2.3434$$
$$\log \text{prod.} = 4.6412$$
$$\log \sqrt{\text{prod.}} = 2.3206$$

$$\log a = 12.2116 - 10$$
$$\log \sqrt{\text{prod.}} = 2.3206$$
$$\log r = 9.8910 - 10$$
$$\therefore r = \boxed{+.778}$$

TABLE 30. ILLUSTRATING THE COMPUTATION OF THE PRODUCT-MOMENT CORRELATION COEFFICIENT WITH THE DATA IN A FREQUENCY TABLE

CHICAGO \ OTIS	80	90	100	110	120	130	140	150	160	170	180	190	f_y	d_y	$f_y d_y$	$f_y d_y^2$
80								5 1	5 1				2	5	10	50
75											4 1		1	4	4	16
70						3 1	3 1	6 2					4	3	12	36
65						4 2	4 2	4 2	4 2				8	2	16	32
60							3 3	4 4	1 1				8	1	8	8
55					0 1	0 3	0 2	0 2		0 1			9	0	0	—
50				−1 1	−1 1	−4 4	−1 1						7	−1	−7	7
45				−4 2	−6 3	−2 1							6	−2	−12	24
40		−3 1	−6 2	−3 1									4	−3	−12	36
35	−4 1												1	−4	−4	16
30															15	225
f_x	1	1	2	4	5	11	9	11	4	1	1		50		$\Sigma f_y d_y$	$\Sigma f_y d_y^2$
d_x	−5	−4	−3	−2	−1	0	1	2	3	4	5					
$f_x d_x$	−5	−4	−6	−8	−5	0	9	22	12	4	5		24	$\Sigma f_x d_x$		
$f_x d_x^2$	25	16	18	16	5	—	9	44	36	16	25		210	$\Sigma f_x d_x^2$		
$\Sigma' f_{xy} d_y$	−4	−3	−6	−8	−7	+1	9	19	10	0	4		15	$\Sigma f_y d_y$		
$d_x(\Sigma' f_{xy} d_y)$	20	12	18	16	7	0	9	38	30	0	20		170	$\Sigma f_{xy} d_x d_y$		

Check

The computation down to the quantities $\Sigma f_x d_x^2$ and $\Sigma f_y d_y^2$ is the same as for the standard deviation, so that the values for b and c may be readily obtained.

The calculation for a presents a little more difficulty. The quantity $\Sigma f_{xy} d_x d_y$ is the result of multiplying each cell frequency by its d_x and d_y and then adding all the products so formed. A more convenient method of calculation, however, is to multiply the cell frequencies in a particular column by the appropriate d_y values, add the results thus found, and multiply this sum by the d_x value of the column. Continuing in this way for all the columns, and adding the products thus found gives the required $\Sigma f_{xy} d_x d_y$. Thus, the frequencies in the column at 150–160 on Otis have been multiplied by the corresponding d_y values and the results recorded in the lower left corners of the cells as 5, 6, 4, 4, 0 (coming down from the top). The sum of these quantities is 19, which, when multiplied by the d_x value 2, gives 38 as the contribution of this column to the total $\Sigma f_{xy} d_x d_y$. The same result would have been obtained if the cell frequencies had been multiplied by d_x and d_y at the same time and added, thus:

$$1\times2\times5+2\times2\times3+2\times2\times2+4\times2\times1+2\times2\times0$$
$$= \quad 10 \quad + \quad 12 \quad + \quad 8 \quad + \quad 8 \quad + \quad 0 \quad = 38$$

The work is shortened, however, by factoring out d_x, which is common to all the products.

The symbol Σ has been used to indicate summation over the whole table, that is, over N items. In order to distinguish summation over the arrays (columns or rows), this has been designated in the table by Σ'. Thus, $\Sigma' f_{xy} d_y$ means the sum for one column of f_{xy} multiplied by the corresponding values d_y.

A very useful check on the computation of a is shown by the double arrow in Table 30. The sum of the quantities $\Sigma' f_{xy} d_y$ should be the same as $\Sigma f_y d_y$, or $\Sigma(\Sigma' f_{xy} d_y) = \Sigma f_y d_y$. Until he becomes proficient in working with the numbers in the corners of the cells, the student should always use this check.

The correction $\frac{(\Sigma f_x d)(_x\Sigma f_y d_y)}{N}$ applied to $\Sigma f_{xy} d_x d_y$ will sometimes be positive and sometimes negative, and it should be remembered that it is to be subtracted algebraically. The remainder of the calculation for r is shown in the example both by straight arithmetic and by logarithms which are recommended.

Computations of this length need to be carefully planned and arranged in order that they may be done quickly and accurately. A standard form has therefore been prepared by the writer. The data may be recorded directly on this sheet (Table 31), and the calculations performed very rapidly.

5. Lines of Regression

In the problem just worked out, it was assumed that the trend of the data in the correlation table was such that the linear method might be applied. As already pointed out, the maximum correlation will occur when all the points fall on a straight line; but with any scatter in the data, two lines will be obtained for a correlation table, one from the means of the columns and one from the means of the rows. The graphical test for approximate linearity and the justification of the use of the product-moment method are furnished by plotting the means of these arrays and noting the extent to which they fall on these two straight lines. A more rigorous test will be given in Chapter X, where it is shown that lack of such linearity reduces the amount of the product-moment correlation.

The curves fitting the means of the columns and rows are known as *regression curves*. They will be illustrated with a larger body of data than that used above, on account of the small number of frequencies in the Otis-Chicago table. The material in Table 32 was supplied through the courtesy of Mr. Douglas E. Scates.* The values on the horizontal scale are the percentage

* "A Study of High-School and First-Year University Grades," *The School Review*, Vol. XXXII (March, 1924).

TABLE 31. CORRELATION FORM FOR PRODUCT-MOMENT METHOD

$$r = \frac{\Sigma f_{xy} d_x d_y - \frac{(\Sigma f_x d_x)(\Sigma f_y d_y)}{N}}{\sqrt{\left[\Sigma f_x d_x^2 - \frac{(\Sigma f_x d_x)^2}{N}\right]\left[\Sigma f_y d_y^2 - \frac{(\Sigma f_y d_y)^2}{N}\right]}} = \frac{a}{\sqrt{bc}}$$

$\log b =$

$\log c =$

$\log \text{prod.} =$

$\log \sqrt{p} =$

$\log a =$

$\log \sqrt{p} =$

$\log r =$

$r =$ [• ± •]

$a = \Sigma f_{xy} d_x d_y - \frac{(\Sigma f_x d_x)(\Sigma f_y d_y)}{N} =$ [ ]

$b = \Sigma f_x d_x^2 - \frac{(\Sigma f_x d_x)^2}{N} =$ [ ]

$c = \Sigma f_y d_y^2 - \frac{(\Sigma f_y d_y)^2}{N} =$ [ ]

TABLE 32. CORRELATION TABLE FOR UNIVERSITY AND HIGH-SCHOOL AVERAGES

University Average	High-School Average 80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	Total
A 6																1		1	2
$\frac{2}{3}$								1	1				3	2	2		4	2	15
$\frac{1}{3}$					1				2	1	4	2	3	1	5	1	2	2	24
A– 5			1			1		2	4	5	5	7	2	5	5	6	6	1	50
$\frac{2}{3}$			1					2		6	6	7	5	4	4	8	5	1	49
$\frac{1}{3}$		1		2	2	5	2	4	3	5	9	4	7	4	8	6		1	63
B 4	1	1	2	4	5	8	5	11	7	13	8	12	9	11	7	4	3		111
$\frac{2}{3}$	2	5	3	6	5	11	4	8	10	4	3	6	9	7	5	1			89
$\frac{1}{3}$	3	3	5	9	7	12	13	8	5	16	10	10	9	2	2	2	1		117
B– 3	3	7	11	10	14	11	13	14	10	11	18	9	15	6	2	2			156
$\frac{2}{3}$	1	12	7	14	11	8	24	13	12	10	9	5	6	2			1		135
$\frac{1}{3}$	4	20	17	25	21	18	18	14	14	4	10	8	4	1	1				179
C 2	5	24	17	20	29	29	23	12	11	14	3	7	4	1	2				201
$\frac{2}{3}$	1	26	28	28	17	20	12	7	6	3	3		1						152
$\frac{1}{3}$	3	23	19	15	16	11	3	8	2	6	1	2			1				110
C– 1	6	12	11	16	10	8	8	7	1	1	1	1							82
$\frac{2}{3}$	4	11	11	9	8	7	9	1	3										63
$\frac{1}{3}$	1	7	7	3	7	1	3		1	1									31
D 0		13	5	10	4	3	3												38
$-\frac{1}{3}$	1	3	2	2	1					1		1							11
$-\frac{2}{3}$		1	2				1	1											5
E –1	2	7	5	3	2	4	1												24
Total	37	176	154	176	160	157	142	113	92	101	90	81	77	46	45	30	23	7	1707

grades of students based on four years' work in high school, while the vertical scale gives the average mark of these students in grade points after three quarters of residence in The University of Chicago. If the means of the columns be denoted by $\overline{Y}_x$ and the means of the rows by $\overline{X}_y$, these values may be computed from the table and arranged as follows:

TABLE 33. MEANS OF THE COLUMNS AND ROWS OBTAINED FROM TABLE 32

X	$\overline{Y}_x$	Y	$\overline{X}_y$
80.5	1.75	−1.00	83.0
81.5	1.54	−0.67	84.1
82.5	1.66	−0.33	83.9
83.5	1.87	0.00	83.2
84.5	1.99	0.33	83.7
85.5	2.23	0.67	83.9
86.5	2.22	1.00	84.1
87.5	2.70	1.33	84.3
88.5	2.86	1.67	84.3
89.5	3.08	2.00	85.5
90.5	3.38	2.33	85.6
91.5	3.40	2.67	86.7
92.5	3.58	3.00	87.7
93.5	3.95	3.33	87.9
94.5	4.20	3.67	88.1
95.5	4.37	4.00	89.7
96.5	4.81	4.33	90.6
97.5	5.14	4.67	92.6
		5.00	92.2
		5.33	92.6
		5.67	94.0
		6.00	95.5

In Fig. 40 the means of the columns have been plotted as dots and the means of the rows as crosses. The former array of points fits rather closely a straight line drawn through them, but the means of the rows form an irregular curve. While freehand curves drawn through both sets of points would give rough approximations to the regression curves, it is better to employ a mathematical method for fitting such curves. In the following paragraph we shall, therefore, discuss a method for obtaining the best-fitting regression curve in the form of a straight line.

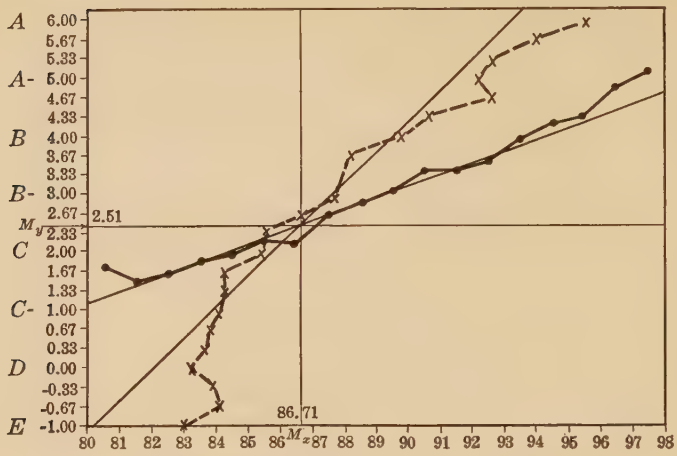


FIG. 40. Regression lines for 1707 grades

The dots represent the means of the columns and the crosses the means of the rows

If we select the line for the means of the columns it is evident that, when taken through the mean of the table at M (Fig. 41), its equation will be of the form

$$\bar{y} = mx,$$

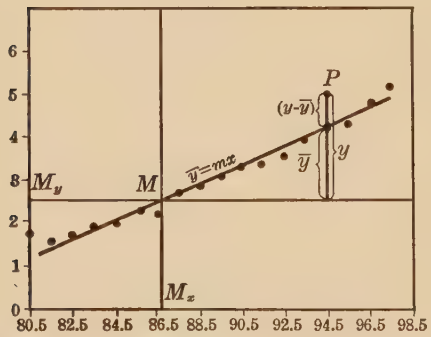


FIG. 41. Illustrating the derivation of the regression line

where m is the slope to be determined, and $\bar{y}$ denotes a point on the line. If P represents any point in the table, its vertical deviation from the line will be $y - \bar{y}$, as shown in the accompanying figure. The problem now is to select m so that the sum of the squares of these deviations (residuals) for the N points shall be as small as possible, that is, so that $\Sigma(y - \bar{y})^2$ shall be a minimum. Replacing $\bar{y}$ by mx, and expanding, we may write the necessary condition in the form

$$\Sigma y^2 - 2\,m\Sigma xy + m^2\Sigma x^2 = \text{a minimum},$$

or
$$\sigma_y^2 - 2\,mr\sigma_x\sigma_y + m^2\sigma_x^2 = \text{a minimum}.$$

Differentiating * this last expression with respect to m, and setting the result equal to zero, gives

$$m = r\frac{\sigma_y}{\sigma_x},$$

which is known as the *regression coefficient of* y *on* x. The required equation through the origin thus becomes

$$\bar{y} = \left(r\frac{\sigma_y}{\sigma_x}\right)x. \quad \left\{\begin{array}{c}\text{Regression line for}\\ \text{means of columns, referred to mean of table}\end{array}\right\} \quad (36)$$

In case the student is not familiar with the differential calculus, the above result may also be shown in the following way:

Setting
$$S_y^2 = \sigma_y^2 - 2\,mr\sigma_x\sigma_y + m^2\sigma_x^2,$$

we shall assume m to have the value $r\frac{\sigma_y}{\sigma_x}$, and show that any different value will produce a larger squared sum. It follows that

$$S_y^2 = \sigma_y^2 - 2\,r^2\sigma_y^2 + r^2\sigma_y^2,$$

or
$$S_y = \sigma_y\sqrt{1 - r^2}. \quad \left\{\begin{array}{c}\text{Standard error}\\ \text{of estimate}\end{array}\right\} \quad (37)$$

Taking
$$m = r\frac{\sigma_y}{\sigma_x} + \delta,$$

we find that

$$S'_y{}^2 = \sigma_y^2 - 2\,r^2\sigma_y^2 - 2\,r\sigma_x\sigma_y\delta + r^2\sigma_y^2 + 2\,r\sigma_x\sigma_y\delta + \sigma_x^2\delta^2,$$

or
$$S'_y{}^2 = \sigma_y^2(1 - r^2) + \sigma_x^2\delta^2.$$

No matter whether δ be positive or negative, $S'_y{}^2$ is greater than S_y^2 and the minimum value for m is therefore $r\frac{\sigma_y}{\sigma_x}$.

By similar reasoning it may be shown that

$$\bar{x} = \left(r\frac{\sigma_x}{\sigma_y}\right)y, \quad \left\{\begin{array}{c}\text{Regression line for}\\ \text{means of rows, referred}\\ \text{to mean of table}\end{array}\right\} \quad (38)$$

* If the reader is unfamiliar with the calculus he should skip to the following paragraph.

where $r\frac{\sigma_x}{\sigma_y}$ is the regression coefficient of x on y. The two regression lines given by (36) and (38) furnish not only the best fit to all the points in the table but also to the means of the columns and rows when the deviations are weighted by the frequencies of the arrays.* The regression coefficients $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$ give the average change in $\bar{y}$ and $\bar{x}$ for a unit change in x and y, respectively.

In case the variables are taken from the origins of the measurements, equations (36) and (38) may be transformed by the relations $x = X - M_x$ and $y = Y - M_y$, giving

$$\bar{Y} = r\frac{\sigma_y}{\sigma_x}X - r\frac{\sigma_y}{\sigma_x}M_x + M_y \quad \left\{\begin{array}{c}\text{Regression}\\ \text{lines in score}\\ \text{form}\end{array}\right\} \quad (39)$$

and

$$\bar{X} = r\frac{\sigma_x}{\sigma_y}Y - r\frac{\sigma_x}{\sigma_y}M_y + M_x. \quad (40)$$

Using the notation a, b, and c as in the calculation of the correlation coefficient, and denoting the class intervals on X and Y by h and k, respectively, two other equations may be obtained. Since $r = \frac{a}{\sqrt{bc}}$, $\sigma_x = \left(\sqrt{\frac{b}{N}}\right)h$, and $\sigma_y = \left(\sqrt{\frac{c}{N}}\right)k$, it follows that

$$\bar{Y} = \frac{ak}{bh}X - \frac{ak}{bh}M_x + M_y \quad \left\{\begin{array}{c}\text{Regression lines in}\\ \text{score form and sym-}\\ \text{bols on correlation}\\ \text{sheet}\end{array}\right\} \quad (41)$$

and

$$\bar{X} = \frac{ah}{ck}Y - \frac{ah}{ck}M_y + M_x. \quad (42)$$

These last equations are the easiest to calculate, since all the necessary quantities may be obtained directly from those given in the work for the correlation coefficient.

For the university and school data in Table 32 we find that

$$a = 17{,}468, \qquad M_x = 86.71,$$
$$b = 28{,}838, \qquad M_y = 2.51,$$
$$c = 28{,}249, \qquad h = 1, \text{ and } k = \tfrac{1}{3}.$$

* Yule, Introduction to Statistics, p. 172.

Substituting these values in equations (41) and (42), we find

that $\bar{Y} = .2019\ X - 15.00$

and $\bar{X} = 1.855\ Y + 82.05.$

These regression lines are plotted in Fig. 40 on page 158.

Representing the regression coefficients by

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{ak}{bh} \qquad (43)$$

{Regression coefficients}

and

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{ah}{ck}, \qquad (44)$$

it follows at once that $r = \sqrt{b_{xy} \cdot b_{yx}}$. For the above data, therefore, $r = \sqrt{.3745245} = .6120$. The correlation coefficient is thus the geometric mean of the two regression coefficients. Equations (43) and (44) also show that while b_{yx} and b_{xy} are functions of scale units, their product is a pure number.

Returning to the equation for the means of the columns, it is evident that it may prove useful in predicting the most probable (average) university grades $\bar{Y}$ for given high-school grades X. Thus a student entering The University of Chicago with a high-school average of 95 will probably make a university average of $.2019 \times 95 - 15.00 = 4.18$ grade points, or a little better than B; while a student entering with a high-school average of 85 will most likely have a university average of 2.16, or slightly better than the required C.

A measure of the value of these predictions is given by the standard deviation of all the observed variations from the regression line. This quantity, which is known as the *standard error of estimate*, has already been presented in equation (37) for the line through the means of the columns. Working out a similar formula for the rows and multiplying the results by .6745 in order to obtain the probable error * of estimate, we have

* For a complete discussion of probable error, see Chapter XIII. The probable error is so defined that half of the errors lie within the limits, mean − probable error and mean + probable error, or $M \pm P.E.$

$$P.E.\,(\text{est. } Y) = .6745\,\sigma_y\sqrt{1-r^2} \quad \left\{\begin{matrix}\text{Probable error of estimate}\\ \text{in predicting } Y \text{ from } X\end{matrix}\right\} \quad (45)$$

and

$$P.E.\,(\text{est. } X) = .6745\,\sigma_x\sqrt{1-r^2}. \quad \left\{\begin{matrix}\text{Probable error of estimate}\\ \text{in predicting } X \text{ from } Y\end{matrix}\right\} \quad (46)$$

Since $\sigma_y = \left(\sqrt{\frac{c}{N}}\right)k = 1.356$, the probable error of estimate for university grades is $.6745(1.356)\sqrt{1-(.612)^2} = .723$. Such calculations are facilitated by the use of logarithms of $\sqrt{1-r^2}$ given on page 54 of Holzinger's Tables.

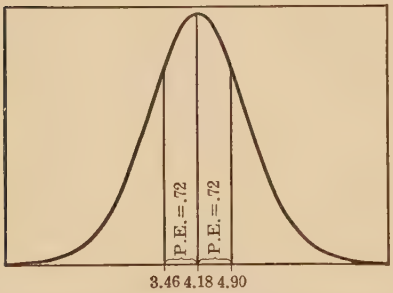


FIG. 42. Illustrating the probable error of estimate

The complete equation for prediction may now be written.

$$\overline{Y} = .2019\,X - 15.00 \pm .72.$$

The use of this equation may be illustrated in the case of a student with a high-school average of 95. Substituting $X = 95$, we find that $\overline{Y}$, or the most probable university average, is 4.18 grade points. This result is written $4.18 \pm .72$, with the interpretation that it is an even chance that the student's university average will be anywhere from 3.46 to 4.90 grade points, or between B− and A−. This conclusion may be drawn from the fact that half of the observed deviations from the predicted mean (4.18) lie between these two values, or, as shown in Fig. 42, half the area of the curve lies between these limits. (See Chapter XII for further interpretation of probable error.)

As might be expected, the value of the prediction becomes better as the correlation increases. When $r = 1$, the standard error of estimate is zero and the prediction is perfect in the sense that all the observed values lie on a straight line. For a list of cautions to be observed in using regression equations, see Chapter XV, section 4.

The meaning of the term "regression," which is due to Sir Francis Galton, may be shown in the case of the line for predicting the height of sons from the height of their fathers. The equation of this line is approximately

$$S = .5\,F + 34'' \qquad (M_S = M_F = 68'')$$

where S represents the son's height and F the father's height in inches. By substituting a few values of F near the mean we find the results which are listed in the following table.

	S	F	$S - F$
	64	60	+4
	65	62	+3
	66	64	+2
	67	66	+1
Mean	68	68	0
	69	70	−1
	70	72	−2
	71	74	−3
	72	76	−4

The column $S - F$ shows regression, or the tendency of the son's predicted height to be nearer the mean than the height of his father. Thus a father 74 inches tall will be expected to have a son only 71 inches in height, while a father 62 inches tall will most probably have a son 65 inches in height, the son's height each time regressing toward the mean of the race. This is one of the important laws of inheritance.

Galton's term "regression" continues to be used for any sort of curve which fits the means of the arrays in a correlation table, even though no problem of inheritance is to be considered.

6. The Interpretation of the Correlation Coefficient

In the case of perfect correlation between two variables the association is regarded by some writers as a causal one, and a smaller degree of correlation as an approximation to causal relationship. It is usually best, however, to avoid this inter-

pretation and to regard the correlation coefficient merely as a mathematical expression for the degree of association between the traits, regardless of the factors producing the result. This may be illustrated by the correlation between height and score on an intelligence test with a group of pupils from Grades III to VII. The correlation found was .71, but it would be absurd to say that the physical growth caused the mental growth, or vice versa. The relationship observed was largely due to a third factor, age, which when eliminated reduced the amount of the correlation to only — .06.

All statistical data are affected by a multiplicity of factors which may obscure the meaning of the relationship found between two observed variables. For example, the correlation between high-school and university grades found on page 161 was .612, a result doubtless due in a large measure to the mentality of the student. Many other factors, however, such as his age, sex, nationality, health, ambition, methods of study, regularity of attendance, and personal appearance, as well as the type of examinations and reaction of the instructors, doubtless contribute also to the observed correlation. Scholarship as measured by marks is thus a variable made up of a large number of other variables, and the correlation found is of doubtful meaning so far as causes are concerned; its main value here is for predicting university grades from high-school grades regardless of the factors affecting such estimates.

With standardized tests it is possible to obtain results which give a better approximation to the correlation between simple variables. The tests themselves measure fairly well certain aspects of human abilities such as rate in reading, accuracy in arithmetic, and quality in handwriting. Proper methods of administering and scoring the tests will eliminate to a large extent errors of the observer, while such outstanding factors as age, sex, grade, and nationality may be controlled by selection of the cases. The correlation between rate and comprehension in reading on a certain test for fifty pupils aged twelve and in

the seventh grade has therefore a good deal more meaning than the correlation for scholarship quoted above.

One rather common interpretation of correlation is that it shows the percentage of agreement between the associated traits. Thus a coefficient of .90 would show 90 per cent agreement, while a coefficient of .45 would show 45 per cent agreement. This interpretation is entirely misleading since the intensity of association does not vary directly as the size of the correlation coefficient.

Another custom in dealing with correlation is to classify the coefficients as "high," "medium," or "low." Thus .75 would generally be regarded as "high," while .25 would be considered as "low." This terminology may be convenient in dealing with test material where the percentage of coefficients above .75 and below .25 is small, but may be misleading when dealing with other types of data. In an age-grade table, for example, a correlation of .75 would be found by comparison with similar coefficients to be relatively low. Another misconception sometimes occurs in interpreting a "high" coefficient, such as .7, as meaning almost perfect agreement. How far this is from the truth may be seen by mere inspection of the scatter diagram for values of this size.

An interpretation of the correlation coefficient that is of some theoretical interest may be illustrated by a problem in dice throwing known as Weldon's experiment.* Twelve dice were shaken in a box and thrown again and again, the number of dice showing four or more spots on the upper face being recorded. When the results of the first, third, fifth, etc. throws were paired against the results of the second, fourth, sixth, etc. throws, no correlation was found because all the events were quite independent of one another.

Next, half of the dice were stained red and after throwing them all and counting all those showing four or more spots,

* William Brown, Essentials of Mental Measurement, p. 78. Cambridge University Press, England, 1911.

the second and every alternate throw thereafter was made by leaving the red dice upon the table, but counting both colors when computing the score. The number due to the red dice was thus common to the two scores. By continuing in this way, two series of odd and even throws were formed for which the correlation approached the value 6/12, or .5. In more general terms, if n is the total number of dice thrown, and c the number common to the pairs of throws, the expected correlation will be

$$r = \frac{c}{n}.*$$

The correlation coefficient may thus be regarded as the ratio of the number of equally effective elements which two variables have in common to the total number of independent elements constituting each, or, more briefly, as the proportion of common elements or causes. It is hardly necessary to add that this interpretation is little more than suggestive in dealing with ordinary statistical data where systems of causation are extremely complicated.

A final interpretation of correlation arises from a consideration of the standard error of estimate, $\sigma_y\sqrt{1-r^2}$. This quantity, as already noted, gives the error in prediction by use of a single score with the regression equation $\bar{y} = r\frac{\sigma_y}{\sigma_x}x$. In case $r = 0$, the prediction has a standard error which is equal to the standard deviation of the predicted variable and is therefore no better than that which would be obtained by selecting a value of y at random from the observed distribution. As r becomes larger, however, the predictive value of a single score becomes better than that afforded by such a chance estimate, the improvement being measured in percentage terms by

$$I_p = 100\left(\frac{\sigma - \sigma\sqrt{1-r^2}}{\sigma}\right) = 100\left(1-\sqrt{1-r^2}\right). \left\{\begin{array}{l}\text{Improvement}\\ \text{over chance in}\\ \text{prediction by}\\ \text{a single score}\end{array}\right\} \quad (47)$$

* For proof see William Brown, Essentials of Mental Measurement, p. 79.

As an example, it may be noted that a correlation coefficient of .80 gives a value of $I_p = 40$, which means that the regression forecast with a single score is here only 40 per cent better than a random guess.

Interpretations of this sort may be useful in dealing with correlations as a basis for linear prediction, but they are likely to give a false impression of the intensity of association given by the correlation coefficient itself. Thus a coefficient of .80 may be rather unsatisfactory for estimates by a single score, but the value $I_p = 40$ can hardly lead us to regard .80 as a "poor" correlation with test scores, as one writer suggests.

It should also be noted that the interpretation afforded by formula (47) is, like the preceding attempts, quite arbitrary. A very different but equally valid percentage measure could be easily derived by considering the squared error of estimate instead of the first power.

The writer is of the opinion that "layman's interpretations" of correlation coefficients should ordinarily be avoided. Such attempts, as already pointed out, are usually based on arbitrary assumptions and may furnish quite misleading and inconsistent results in the hands of the "layman." The student of statistics will soon find that he needs no such devices and his best and most useful guide in interpreting correlations will be given by a simple scatter diagram with fitted regression curves. Comparisons between correlation coefficients should always be made through the use of the sampling errors discussed in Chapter XIII.

7. Some Uses of Correlation in Evaluating Test Material

In the preparation of standardized test material several terms have come to be accepted with very definite meanings. One of these is the *reliability* of the test, which may be defined as the consistency with which the test measures what it purports to measure. An index of this consistency is given by the *reliability*

coefficient, which is merely the correlation between two forms of the same test given at different times. If X_1 and X_I be the two test forms, this coefficient may be expressed as r_{1I}.

Another characteristic of a test is its *validity*,* by which is meant the extent to which it does measure what it purports to measure. The evidence in this case must of course be indirect. It is customary to select some criterion C, which is known to index the trait in question. The correlation r_{cx} between the criterion and the test therefore furnishes numerical evidence of the validity of the latter.

Suppose, for example, that five tests are proposed as measures of intelligence. By correlating the results of each of these with the scores on some accepted scale such as the Stanford-Binet, a series of validity coefficients of the form $r_{c1} = .78$, $r_{c2} = .82$, $r_{c3} = .40$, $r_{c4} = .78$, and $r_{c5} = .43$ might be obtained. Tests X_1, X_2, and X_4 would thus be regarded as considerably more valid than the other two. In case it is objected that high correlation is no sure evidence that the tests are measuring the same thing as the criterion, it may be argued that this is the best evidence we have and that such correlation shows that the tests have high predictive value, which is sufficient justification for their use.

In case a number of similar tests are pooled the reliability and validity coefficients of the lengthened test may be obtained by applying Professor Spearman's † theorem on the correlation of sums and differences. These new formulas will be derived directly, however, making use of *standard scores* such as $z = \frac{x}{\sigma}$ (see Chapter VII).

Let $z_1 = \frac{x_1}{\sigma_1}$ and $z'_1 = \frac{x'_1}{\sigma'_1}$ be the standard scores on two similar tests, and let $z_I = \frac{x_I}{\sigma_I}$ and $z'_I = \frac{x'_I}{\sigma'_I}$ be the standard scores on

* Instead of using the term "validity" some test workers prefer to speak of the "*predictive* value" of a test. This is essentially the same property as validity, inasmuch as both are measured by correlating the test with some criterion.

† C. Spearman, "Correlations of Sums and Differences," *British Journal of Psychology*, Vol. V (1913), p. 417.

comparable forms of each. The problem is to determine the reliability of $z_1 + z'_1$, knowing the reliability of z_1. This will be given by working out

$$r_{(z_1+z'_1)(z_I+z'_I)} = \frac{\Sigma z_1 z_I + \Sigma z_1 z'_I + \Sigma z'_1 z_I + \Sigma z'_1 z'_I}{N\sigma_{(z_1+z'_1)}\sigma_{(z_I+z'_I)}},$$

which comes from expanding $r = \dfrac{\Sigma xy}{N\sigma_x\sigma_y}$ when $x = z_1 + z'_1$ and $y = z_I + z'_I$. The standard deviations $\sigma_{(z_1+z'_1)}$ and $\sigma_{(z_I+z'_I)}$ reduce to $\sqrt{2+2r_{1I}}$, since $\sigma_{z_1} = \sigma_{z'_1} = \sigma_{z_I} = \sigma_{z'_I} = 1$. All the correlations between the z's are equal to r_{1I}. We therefore have

$$r_{(z_1+z'_1)(z_I+z'_I)} = \frac{2\,r_{1I}}{1+r_{1I}}.$$

By induction it may be easily shown that by increasing a test n-fold with similar material the cumulative reliability coefficient is given by

$$r_{nn} = \frac{nr_{1I}}{1+(n-1)r_{1I}}. \left\{\begin{array}{c}\text{Spearman-Brown formula for predicting}\\ \text{reliability of lengthened tests}\end{array}\right\} \quad (48)$$

This expression is often called the Spearman-Brown prophecy formula, since it was proved independently by both men.

As an example let us assume that a test with a reliability coefficient of .7 has been prepared. What will the reliability be when the test has been made three times as long by the addition of similar material? The answer is found by substituting $n = 3$ and $r_{1I} = .7$ in equation (48), giving

$$\frac{3 \times .7}{1+2 \times .7} = .875.$$

An empirical check on this formula was made by Miss Blythe Clayton* and the writer, with carefully graduated spelling material. Seven equally difficult tests with parallel forms were given and the results of actual pooling compared with those

* Karl J. Holzinger and Blythe Clayton, "Further Experiments in the Application of Spearman's Prophecy Formula," *Journal of Educational Psychology* (May, 1925), Vol. XVI, pp. 289–299.

predicted by the formula. The close agreement between observed and theoretical values is shown in the following table.

TABLE 34. OBSERVED AND PREDICTED RELIABILITY COEFFICIENTS FOR SUCCESSIVE POOLS OF n EQUALLY DIFFICULT SPELLING TESTS

NUMBER OF TESTS POOLED = n	OBSERVED RELIABILITY COEFFICIENT	THEORETICAL COEFFICIENT FROM FORMULA (48)
1	.743	.743
2	.841	.853
3	.906	.897
4	.916	.920
5	.941	.936
6	.949	.945
7	.955	.953

The formula for the validity of n pooled tests may be obtained by working out

$$r_{c(z_1 + z_2 + z_3 + \cdots + z_n)}.$$

For three such tests we shall have

$$r_{c(z_1 + z_2 + z_3)} = \frac{\Sigma cz_1 + \Sigma cz_2 + \Sigma cz_3}{N\sigma_c \sigma_{(z_1 + z_2 + z_3)}}.$$

Substituting the values for Σcz and $\sigma_{(z_1 + z_2 + z_3)}$, there results

$$r_{c(z_1 + z_2 + z_3)} = \frac{r_{cz_1} + r_{cz_2} + r_{cz_3}}{\sqrt{3 + 2\,r_{z_1z_2} + 2\,r_{z_1z_3} + 2\,r_{z_2z_3}}}, \qquad (49)$$

or

$$r_{c(z_1 + z_2 + z_3)} = \frac{3\,r_{cz}}{\sqrt{3 + 6\,r_{zz}}}, \qquad (50)$$

if the validity coefficients and correlations r_{zz} are equal. By induction it may now be shown that for n tests we have

$$r_{cn} = \frac{nr_{cz}}{\sqrt{n + n(n-1)r_{zz}}}. \qquad \left\{\begin{array}{c}\text{Formula for predicting validity}\\ \text{of lengthened tests}\end{array}\right\} \qquad (51)$$

A test with a reliability of .7 and a validity coefficient of .6 would, upon being made three times as long, have a validity coefficient of

$$\frac{3 \times .6}{\sqrt{3 + 6 \times .7}} = .67.$$

An interesting agreement between psychological theory and statistical analysis may be noted in connection with formula (49). Suppose several tests are to be pooled for the measurement of intelligence. Some psychologists would select tests which correlate high with a criterion and low amongst themselves so as to obtain not only those which are most valid but also those which measure as wide a sampling of intellectual abilities as possible. Psychological theory would then require a pool of tests with high values r_{cz_1}, r_{cz_2}, r_{cz_3}, etc., and low values $r_{z_1z_2}$, etc. Very fortunately, such a combination will produce a high validity coefficient for the combined tests, as may be seen from formula (49). High coefficients, r_{cz}, will give a large numerator, and low coefficients, r_{zz}, will give a small denominator, both acting in the same direction to produce a high validity coefficient for the pooled tests.

Another interesting application of correlation occurs in connection with the scoring of multiple-choice tests. The general formula advocated to correct for the element of guessing is

$$S = R - \frac{1}{(n-1)}W = R - CW, \quad \left\{\begin{matrix}\text{Multiple-response} \\ \text{scoring formula}\end{matrix}\right\} \qquad (52)$$

where S is the score, R the number of right responses, W the number of wrong responses, C a constant, and n the number of choices. Thus if the examinee is to underline one of three suggested answers, he would be scored by the formula $S = R - \frac{1}{2}W$.

Such complicated scoring methods may be avoided entirely if all pupils be allowed to finish the test. In this case, if $A =$ the number of attempts, $R + W = A =$ a constant. We may also write $S = R + C(R - A)$, or $S = aR + b$ where a and b are constants. The correlation between S and R will now be perfect, so that the number of "rights" furnishes as reliable and valid a score as the full formula. The proof that $r_{SR} = +1.00$ is left as an exercise for the student. (See Exercise 8.)

8. The Effect of Selection upon Correlation

If the correlation between two traits, X_1 and X_2, is given by r_{12} with a sample of N and selected values are chosen, reducing the size of the sample to $N - n$, then the resulting correlation R_{12} will differ from that obtained for the unselected group.

Professor Pearson* has shown that if σ_1 denotes the variability in X_1 before selection and Σ_1 the variability after selection, then

$$R_{12} = \frac{\Sigma_1}{\sigma_1} \frac{r_{12}}{\sqrt{1 - r_{12}^2 + r_{12}^2 \left(\frac{\Sigma_1}{\sigma_1}\right)^2}} \cdot \left\{\begin{matrix}\text{Correlation} \\ \text{after selec-} \\ \text{tion}\end{matrix}\right\} \quad (53)$$

The correlation R_{12} decreases with Σ_1, so that restricting the range of X_1 lowers the original correlation.

As an example, let us assume that the correlation between two traits is given by $r_{12} = .7$, and that values of X_1 are taken so that $\sigma_1 = 10$ is reduced to $\Sigma_1 = 5$. Substituting these values in equation (53), we find $R_{12} = .44$, which is considerably less than the correlation before selection.

In case there is selection in both variables the adjustment formulas† become very complicated. The beginning student will do well to avoid problems involving such correction until he is in a position to read the papers cited in footnotes below.

9. The Effect of Range of Talent upon Correlation

The magnitude of correlation coefficients clearly depends upon the particular group studied. Thus, "to secure a reliability coefficient of .40 from a group composed of children in a single grade is probably indicative of greater, not less, reliability than to secure a reliability coefficient of .90 from a group com-

* Karl Pearson, "On the Influence of Natural Selection on the Variability and Correlation of Organs," *Philosophical Transactions of the Royal Society of London*, Series A., Vol. CC, p. 23.

† Karl Pearson, "On the Influence of Double Selection on Variation and Correlation of Two Characters," *Biometrika*, Vol. VI (1908).

posed of children from the second to the twelfth grades," as shown by Professor Kelley.*

This difference in the value of the obtained correlation is due to what Kelley calls "range of talent," and he has given a formula (see equation (116)) for adjusting coefficients for varying ranges of talent. The proof of the formula, however, is open to some objections and it is probably better, therefore, to compare correlation coefficients only when they have been obtained from the same group or from groups varying but slightly in range of talent.

As a general caution it may be noted that it is not safe to compare correlation coefficients of any sort obtained from groups where the range of talent or other conditioning factors such as range in age are very different (see Chapter XV).

EXERCISES

1. Make correlation tables for Otis with Terman and for Chicago with Terman tests from the data of Exercise 1, Chapter II. Use intervals of 69.5–79.5 etc. for Otis and Terman, and 29.75–34.75 etc. for Chicago. Work out the coefficients of correlation.

($r_{OT} = .718$; $r_{CT} = .681$. *Ans.*)

2. Make a correlation table for the two spelling tests of Exercise 6, Chapter II, using intervals of 5 units for both tests. Work out the correlation coefficient. ($r = .963$. *Ans.*)

3. Compute the means of the columns and the means of the rows from the table of Exercise 2, and plot them on graph paper. Calculate the equations of the two regression lines and plot on the same graph. Determine the two probable errors of estimate.

($\overline{A} = 1.01B - 2.28 \pm 4.04$; $\overline{B} = .92A + 6.63 \pm 3.85$. *Ans.*)

4. Calculate the correlation coefficient, regression lines, and probable errors of estimate for the table on page 174. Compute the means of the columns and rows and plot with the regression lines as in Exercise 3. The values of the constants are given below the table.

5. Compute the correlation coefficient for the table on page 175.

* T. L. Kelley, "The Reliability of Test Scores," *Journal of Educational Research*, May, 1921, p. 374.

CORRELATION OF BETA RAW SCORE (SUGGESTED FORM) WITH ALPHA RAW SCORE — ENGLISH-SPEAKING WHITES FROM NINE CAMPS*

ALPHA RAW SCORE	BETA RAW SCORE (SUGGESTED FORM)																							NUMBER OF CASES
	0–4	5–9	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99	100–104	105–109	110–114	
180–189	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	1	1	1	3
170–179	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	1	..	..	..	..	1
160–169	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	1	1	3	1	1	7
150–159	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	1	2	2	..	..	5
140–149	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	2	..	4	5	..	..	1	12
130–139	..	..	..	..	..	..	..	..	..	..	..	..	1	..	..	1	2	2	2	2	4	2	1	17
120–129	..	..	..	..	..	..	..	..	..	..	..	1	..	1	1	1	1	5	4	3	2	2	..	21
110–119	..	..	..	..	..	..	..	..	..	..	..	..	1	..	..	3	4	3	2	5	1	1	1	21
100–109	..	..	..	..	..	..	..	..	..	..	..	1	..	2	..	5	1	6	3	2	2	..	..	22
90–99	..	..	..	..	..	..	..	..	..	..	1	..	..	3	3	3	11	8	7	..	3	..	..	39
80–89	..	..	..	..	..	..	..	1	..	..	3	2	4	4	5	14	8	4	1	4	2	..	1	53
70–79	..	..	..	..	..	1	..	1	..	..	1	3	3	2	5	2	7	8	3	2	..	..	..	38
60–69	..	..	..	..	..	..	..	..	..	..	..	4	4	8	7	8	10	7	3	2	..	..	..	53
50–59	..	..	..	..	..	1	..	1	5	2	4	2	4	3	4	5	4	3	..	..	..	..	..	38
40–49	..	..	..	..	..	3	..	2	2	3	5	5	8	3	10	5	5	2	..	..	..	..	..	53
30–39	..	..	1	..	1	2	3	5	9	3	9	7	7	7	7	6	1	..	..	..	..	..	..	68
20–29	..	1	1	..	2	3	5	5	8	6	9	6	8	1	4	1	..	..	..	..	..	..	..	60
10–19	..	..	3	4	7	5	5	10	5	10	6	..	3	2	..	..	..	..	..	..	..	..	..	60
0–9	5	7	10	9	13	12	5	7	8	3	..	1	2	..	..	..	..	..	..	..	..	..	..	82
Number of cases	5	8	15	13	23	27	18	32	37	27	38	32	45	36	46	54	56	48	32	28	20	7	6	653

$r = .806$; $\sigma_\alpha = 42.0$; $M_\alpha = 58.5$; $\sigma_\beta = 25.7$; $M_\beta = 62.7$. *Ans.*

* From Memoirs of the National Academy of Sciences, Vol. XV, p. 392.

CORRELATION BETWEEN ALPHA RAW TOTAL AND REPORTED SCHOOLING. GROUP X: NATIVE-BORN DRAFT, NINE CAMPS*

ALPHA RAW TOTAL SCORE	YEARS SCHOOLING																	TOTAL
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
180–189	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	2	1	3
170–179	..	..	..	..	..	..	..	..	..	..	..	..	..	..	1	..	..	1
160–169	..	..	..	..	..	..	..	1	..	..	..	..	3	1	..	1	1	7
150–159	..	..	..	..	..	..	..	..	..	..	..	2	2	..	..	..	1	5
140–149	..	..	..	..	..	..	..	1	1	1	2	2	1	..	2	1	1	12
130–139	..	..	..	..	..	..	1	..	1	2	4	2	3	1	1	..	2	17
120–129	..	..	..	..	..	..	..	..	5	5	3	3	1	2	1	1	..	21
110–119	..	..	..	..	..	..	..	2	7	1	4	3	4	..	..	..	..	21
100–109	..	..	..	..	..	..	2	2	5	2	3	2	3	2	1	..	..	22
90–99	..	..	..	..	..	1	1	3	11	7	4	3	4	5	..	..	..	39
80–89	..	..	..	..	..	1	3	5	25	7	7	1	1	2	1	..	..	53
70–79	..	..	..	..	..	2	5	9	11	4	1	2	1	1	1	1	..	38
60–69	..	..	..	..	4	2	3	12	15	6	5	4	..	2	..	..	..	53
50–59	..	..	..	..	2	8	6	6	11	2	1	1	1	..	..	..	..	38
40–49	..	..	..	..	2	9	7	14	18	1	1	1	..	..	..	..	..	53
30–39	..	..	..	3	12	8	11	9	21	2	..	..	2	..	..	..	..	68
20–29	..	..	..	5	11	15	7	8	11	..	1	..	1	..	1	..	..	60
10–19	1	2	3	6	18	15	6	2	6	..	..	1	..	..	..	..	..	60
0–9	14	15	4	15	14	16	2	2	..	..	..	..	..	..	..	..	..	82
Total	15	17	7	29	63	77	54	76	148	40	36	27	27	16	9	6	6	653

$r = + 0.752$. *Ans.*

* From Memoirs of the National Academy of Sciences, Vol. XV, p. 780.

6. Work out the correlation coefficient for the first 25 pairs of scores for the data of Exercise 2, using the method described in section 3. Compare the amount of arithmetic with that involved in the use of the correlation table.

7. The experimental reliability coefficients found by lengthening a spelling test from one to ten times the original value were: .850, .903, .927, .946, .960, .970, .974, .976, .980, and .981. Calculate the corresponding theoretical coefficients from the Spearman-Brown formula, using $r_{1t} = .847$ and $n = 1, 2, 3 \cdots 10$ successively. (Data furnished by Professor G. M. Ruch.)

(.847, .917, .943, .957, .965, .971, .975, .978, .980, .982. *Ans.*)

8. Work out the proof for the exercise suggested at the end of section 7.

9. Prove that the correlation between $aX + b$ and $cY + d$ is the same as the correlation between X and Y where a, b, c, and d are constants.

CHAPTER X

NON-LINEAR CORRELATION

1. The Correlation Ratio

As pointed out in Chapter IX, when the means of the arrays do not lie fairly closely on a straight line, the regression is to be regarded as non-linear. The correlation coefficient, which measures the approach to functionality only when the traits have a linear relationship, will give an understatement of the degree of association present for such curvilinear trends and is therefore inapplicable. An extreme case of this understatement is illustrated in Fig. 43, where all the observations lie on a half circle. The correlation as defined by approach to functionality will be perfect, but the product-moment coefficient will give zero as the amount of association. This may be readily verified by noting that from the symmetry of the points, Σxy will equal zero.

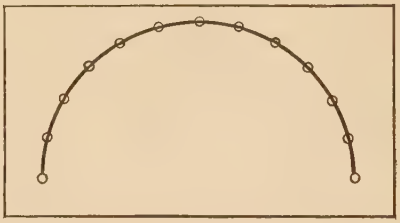


FIG. 43. An extreme case of non-linear correlation

In order to measure the correlation for non-linear tables, Professor Pearson has devised a coefficient known as the *correlation ratio.* The meaning of this coefficient may be shown by returning to formula (37) for the standard error of estimate of y on x. Rearranging the terms in this formula, we have

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}, \quad \left\{\begin{matrix}\text{Correlation coeffi-}\\ \text{cient in ratio form}\end{matrix}\right\} \qquad (54)$$

where S_y is the standard deviation of the differences $y - \bar{y}$, or residuals from estimation by the regression line $\bar{y} = mx$. The

coefficient of correlation was derived on the assumption that the means of the columns, $\bar{y}_x$, and the means of the rows, $\bar{x}_y$, lie on their respective regression lines.

In case the means of the columns, $\bar{y}_x$, do not lie on a straight line, the residuals $y - \bar{y}$ may be replaced by the differences $y - \bar{y}_x$, whose standard deviation is denoted by σ_{ay}. The correlation ratio for the means of the columns may then be defined as

$$\eta_{yx} = \sqrt{1 - \frac{\sigma_{ay}^2}{\sigma_y^2}}, \tag{55}$$

and, for the means of the rows, as

{Correlation ratios, original form}

$$\eta_{xy} = \sqrt{1 - \frac{\sigma_{ax}^2}{\sigma_x^2}}. \tag{56}$$

From Fig. 44 it is apparent that the differences $y - \bar{y}_x$ and their standard deviation σ_{ay} measure the extent to which the points in the scatter diagram are concentrated about the irregular regression curve. When all the points in the diagram are located at the column means, the differences $y - \bar{y}_x$ and σ_{ay} will be zero, giving $\eta_{yx} = 1$; but when there is any scatter in the arrays, σ_{ay} will not be zero, and η_{yx} will be less than 1. The correlation ratio thus measures the approach of the data to any single-valued function, while the correlation coefficient indicates the closeness to linear functionality.

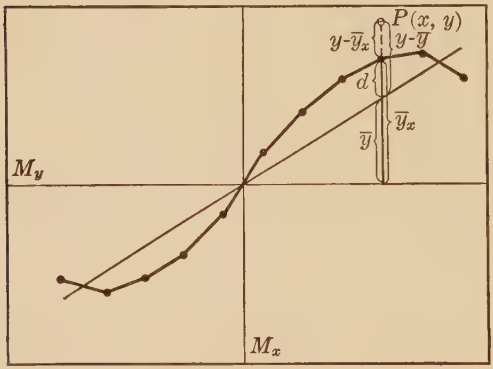


FIG. 44. Illustrating the correlation ratio

It is further evident that if the regression is linear, $y - \bar{y} = y - \bar{y}_x$ for all the columns, so that $S_y = \sigma_{ay}$, and $r = \eta_{yx}$. If the

regression is not linear, $\bar{y}_x = \bar{y} \pm d$, or $y - \bar{y}_x \pm d = y - \bar{y}$. Squaring both members of this last expression, summing over the whole table, and introducing f_x as a symbol of operation, we have

$$\Sigma f_x(y - \bar{y}_x)^2 \pm 2\,\Sigma f_x(y - \bar{y}_x)d + \Sigma f_x d^2 = \Sigma f_x(y - \bar{y})^2.$$

Since the middle term on the left is zero for each column, this reduces to

$$\sigma_{ay}^2 + \sigma_d^2 = S_y^2. \tag{57}$$

By combining equations (54), (55), and (57), we finally obtain

$$\sigma_y^2(\eta_{yx}^2 - r^2) = \sigma_d^2. \left\{\begin{matrix}\text{Relation between correla-}\\ \text{tion coefficient and ratio}\end{matrix}\right\} \tag{58}$$

This proves that η_{yx} is always greater than or equal to r, since σ_d is a positive quantity. The same reasoning might of course be applied to η_{xy}.

From the above discussion it is apparent that the single measure of association furnished by the correlation coefficient may be replaced by the two correlation ratios which are always numerically greater than r. The correlation coefficient fails to measure the full amount of association in case the regression is not linear, and should not be used unless the departure from linearity is negligible (see section 4).

2. Modified Formulas for the Correlation Ratios

Formulas which are more convenient for computation may be obtained by modifying equations (55) and (56) and introducing the methods and notation of Chapter IX. The quantity $y - \bar{y}_x$, when squared and summed over a column, gives

$$\Sigma'(y - \bar{y}_x)^2 = \Sigma' y^2 - 2\,\Sigma' y\bar{y}_x + \Sigma'\bar{y}_x^2 = \Sigma' y^2 - f_x\bar{y}_x^2,$$

where the primes denote summation over a column. Summing next over the whole table, we find

$$\Sigma\Sigma'(y - \bar{y}_x)^2 = \Sigma\Sigma' y^2 - \Sigma f_x\bar{y}_x^2.$$

Denoting the standard deviation of $\bar{y}_x$ by $\sigma_{\bar{y}_x}$, we then have

$$\sigma_{ay}^2 = \sigma_y^2 - \sigma_{\bar{y}x}^2. \tag{59}$$

If this result is substituted in equation (55), we obtain

$$\eta_{yx} = \frac{\sigma_{\bar{y}x}}{\sigma_y} \quad \left\{\begin{array}{l}\text{Correlation ratios as}\\ \text{quotients of two}\\ \text{standard deviations}\end{array}\right\} \tag{60}$$

and, similarly,

$$\eta_{xy} = \frac{\sigma_{\bar{x}y}}{\sigma_x}. \tag{61}$$

The correlation ratio is thus the quotient of the standard deviation of the means of the arrays divided by the standard deviation of the whole table. It should be noted that in forming $\sigma_{\bar{y}_x}$, the deviations are weighted by the frequencies of the arrays.

We shall next modify formulas (60) and (61) so that the calculations may be carried out with the variables taken from arbitrary origins as in the formulas of Chapter IX.

The first formula may be written

$$\eta_{yx} = \frac{\sqrt{\dfrac{\Sigma f_x(M_y - \bar{Y}_x)^2}{N}}}{\sigma_y}, \quad \left\{\begin{array}{c}\text{Correlation ratio}\\ \text{for means of col-}\\ \text{umns}\end{array}\right\} \tag{62}$$

where

$$M_y = A_y + \frac{(\Sigma f_y d_y)k}{N}$$

and

$$\bar{Y}_x = A_y + \frac{(\Sigma' f_{xy} d_y)k}{f_x}.$$

We therefore have

$$\frac{M_y - \bar{Y}_x}{k} = \frac{\Sigma f_y d_y}{N} - \frac{\Sigma' f_{xy} d_y}{f_x}$$

and

$$\frac{(M_y - \bar{Y}_x)^2}{k^2} = \left(\frac{\Sigma f_y d_y}{N}\right)^2 - 2\left(\frac{\Sigma f_y d_y}{N}\right)\left(\frac{\Sigma' f_{xy} d_y}{f_x}\right) + \left(\frac{\Sigma' f_{xy} d_y}{f_x}\right)^2.$$

Summing over the columns and then over the whole table, we find

$$\frac{\Sigma f_x(M_y - \bar{Y}_x)^2}{k^2} = N\left(\frac{\Sigma f_y d_y}{N}\right)^2 - 2\left(\frac{\Sigma f_y d_y}{N}\right)(\Sigma f_y d_y) + \Sigma\left[\frac{(\Sigma' f_{xy} d_y)^2}{f_x}\right],$$

since $$\Sigma\Sigma' \frac{f_{xy}d_y}{f_x} = \Sigma f_y d_y.$$

Upon collecting terms we may write

$$\frac{\Sigma f_x(M_y - \overline{Y}_x)^2}{k^2} = \Sigma\left[\frac{(\Sigma' f_{xy}d_y)^2}{f_x}\right] - \frac{(\Sigma f_y d_y)^2}{N} = e$$

and, similarly,

$$\frac{\Sigma f_y(M_x - \overline{X}_y)^2}{h^2} = \Sigma\left[\frac{(\Sigma' f_{xy}d_x)^2}{f_y}\right] - \frac{(\Sigma f_x d_x)^2}{N} = d.$$

It has already been shown in Chapter IX that

$$\sigma_x = \left(\sqrt{\frac{b}{N}}\right)h \text{ and } \sigma_y = \left(\sqrt{\frac{c}{N}}\right)k.$$

Substituting these results in formula (62), we have

$$\eta_{yx} = \sqrt{\frac{e}{c}} \quad \left\{\begin{array}{c}\text{Correlation ratios for}\\ \text{correlation blank}\end{array}\right\} \tag{63}$$

and, similarly, $$\eta_{xy} = \sqrt{\frac{d}{b}}. \tag{64}$$

These last formulas are especially convenient when the correlation coefficient and ratios are to be compared as is often necessary. The quantities b and c and the corrections for d and e are obtained in working out the correlation coefficient. The remainder of the computation for d and e may be readily done with the aid of the special correlation sheet shown in Table 35.

3. A Combination Form for the Correlation Coefficient and Ratios

In Table 35 the full computation of the correlation ratios is shown. It will be noted that this form differs from the one shown in Table 31 in that two additional columns and rows for the calculation of the ratios are given.

For the items on the row headed $(\Sigma' f_{xy}d_y)^2$ it is only necessary to square, by means of tables, the quantities $(\Sigma' f_{xy}d_y)$ already found in calculating r. In the last row these squared

items are divided by the corresponding frequencies f_x, the total sum being $\Sigma\left[\frac{(\Sigma' f_{xy} d_y)^2}{f_x}\right]$. The correction to this last quantity is $\frac{(\Sigma f_y d_y)^2}{N}$, also known from previous work. Making the correction, we obtain $e = 10{,}827$, and by a similar calculation for the rows we determine $d = 11{,}817$. The remainder of the computation may be readily done with the aid of logarithms as illustrated on the sheet.

It will be noted that the computation for $\Sigma f_{xy} d_x d_y$ has been checked by working out this quantity from both the columns and rows. Other important checks, such as $\Sigma[\Sigma' f_{xy} d_y] = \Sigma f_y d_y$, should be noted on the sheet and carefully observed in the calculations.

The value for η_{yx} comes out as .6191, while η_{xy} is .6401. The former ratio is in close agreement with the correlation coefficient, $r = .6120$, on account of the linearity of the means of the columns. The coefficient η_{xy}, however, is somewhat larger than r because of the irregular regression curve for the rows.

In case the means of the arrays are required for plotting they may be readily found by use of the formulas

$$\bar{X}_y = A_x + \left(\frac{\Sigma' f_{xy} d_x}{f_y}\right) h \qquad \left\{\begin{array}{l}\text{Means of the}\\ \text{arrays in a cor-}\\ \text{relation table}\end{array}\right\} \quad (65)$$

and

$$\bar{Y}_x = A_y + \left(\frac{\Sigma' f_{xy} d_y}{f_x}\right) k \qquad \left\{\begin{array}{l}\text{Means of the}\\ \text{arrays in a cor-}\\ \text{relation table}\end{array}\right\} \quad (66)$$

where A_x and A_y are the assumed means and h and k the class intervals for the variables X and Y, respectively. The quantities $\Sigma' f_{xy} d_x$ and $\Sigma' f_{xy} d_y$ may be taken directly from the correlation sheet. For the means of the columns, we should thus have

$$\bar{Y}_{80.5} = 2.667 + \frac{-102}{37} \times \frac{1}{3} = 1.75,$$

$$\bar{Y}_{81.5} = 2.667 + \frac{-594}{176} \times \frac{1}{3} = 1.54,$$

$$\bar{Y}_{82.5} = 2.667 + \frac{-463}{154} \times \frac{1}{3} = 1.66, \text{ etc.}$$

TABLE 35. FORM FOR CORR

High-s

UNIVERSITY MARKS IN GRADE POINTS.

	80	81	82	83	84	85	86	87	88	89
A. 6										
A- 5										
B. 4										
B- 3										
C. 2										
C- 1										
D. 0										
E. -1										
f_x	37	176	154	176	160	157	142	113	92	10
d_x	-6	-5	-4	-3	-2	-1	0	1	2	
$f_x d_x$	-222	-880	-616	-528	-320	-157	——	113	184	3(
$f_x d^2_x$	1,332	4,400	2,464	1,584	640	157	——	113	368	9(
$\Sigma' f_{xy} d_y$	-102	-594	-463	-423	-323	-205	-191	11	54	1
$d_x \Sigma' f_{xy} d_y$	612	2,970	1,852	1,269	646	205	——	11	108	3
$(\Sigma' f_{xy} d_y)^2$	10,404	352,836	214,369	178,929	104,329	42,025	36,481	121	2,916	15,6
$\frac{(\Sigma' f_{xy} d_y)^2}{f_x}$	281.2	2,004.8	1392	1,016.6	652.1	267.7	256.9	1.1	31.7	15
$\bar{y}_x$	1.75	1.54	1.66	1.87	1.99	2.23	2.22	2.70	2.86	3.(

$$a = \Sigma f_{xy} d_x d_y - \frac{(\Sigma f_x d_x)(\Sigma f_y d_y)}{N} = \boxed{17{,}468}$$

$$b = \Sigma f_x d^2_x - \frac{(\Sigma f_x d_x)^2}{N} = \boxed{28{,}838}$$

$$c = \Sigma f_y d^2_y - \frac{(\Sigma f_y d_y)^2}{N} = \boxed{28{,}249}$$

$$d = \Sigma\left[\frac{(\Sigma' f_{xy} d_x)^2}{f_y}\right] - \frac{(\Sigma f_x d_x)^2}{N} = \boxed{11{,}817}$$

$$e = \Sigma\left[\frac{(\Sigma' f_{xy} d_y)^2}{f_x}\right] - \frac{(\Sigma f_y d_y)^2}{N} = \boxed{10{,}827}$$

$M_x = 86.71$ $\quad M_y = 2.51$

$$r_{xy} = \frac{a}{\sqrt{bc}}$$

$$\eta_{xy} = \sqrt{\frac{d}{b}}$$

$$\eta_{yx} = \sqrt{\frac{e}{c}}$$

PREPARED BY KARL J. HOLZINGER

Coefficient and Ratios

erage

								f_y	d_y	$f_y d_y$	$f_y d_y^2$	$\Sigma' f_{xy} d_x$	$d_y \Sigma' f_{xy} d_x$	$(\Sigma' f_{xy} d_x)^2$	$\frac{(\Sigma' f_{xy} d_x)^2}{f_y}$	$\bar{X}_y$
9	2 9	3 9	4 9	5 9	6 9	7 9	8									
			10 1 8		10 1 10		2	10	20	200	18	180	324	162.0	95.5
	27 3 18	18 2 14	18 2 16		36 4 40	18 2 22	15	9	135	1,215	113	1,017	12,769	851.3	94.0
2 10	24 3 18	8 1 7	40 5 40	8 1 9	16 2 20	16 2 22	24	8	192	1,536	147	1,176	21,609	900.4	92.6
9 35	14 2 12	35 5 35	35 5 40	42 6 54	42 6 60	7 1 11	50	7	350	2,450	287	2,009	82,369	1,647.4	92.2
2 35	30 5 30	24 4 28	24 4 32	48 8 72	30 5 50	6 1 11	49	6	294	1,764	298	1,788	88,804	1,812.3	92.6
4 20	35 7 42	20 4 28	40 8 64	30 6 54		5 1 11	63	5	315	1,575	260	1,300	67,600	1,073.0	90.6
12 60	36 9 54	44 11 77	28 7 56	16 4 36	12 3 30		111	4	444	1,776	360	1,440	129,600	1,167.6	89.7
6 30	27 9 54	21 7 49	15 5 40	3 1 9			89	3	267	801	146	438	21,316	239.5	88.1
10 50	18 9 54	4 2 14	4 2 16	4 2 18	2 1 10		117	2	234	468	162	324	26,244	224.3	87.9
9 45	15 5 90	6 6 42	2 2 16	2 2 18			156	1	156	156	184	184	33,856	217.0	87.7
5 25	0 6 36	0 2 14			0 1 10		135	0	—	—	22	—	484	3.6	86.7
8 40	-4 4 24	-1 1 7	-1 1 8				179	-1	-179	179	-154	154	23,716	132.5	85.6
7 35	-8 4 24	-2 1 7	-4 2 16				201	-2	-402	804	-195	390	38,025	189.2	85.5
	-3 1 6						152	-3	-456	1,368	-340	1,020	115,600	760.5	84.3
2 10			-4 1 8				110	-4	-440	1,760	-245	980	60,025	545.7	84.3
5 5							82	-5	-410	2,050	-195	975	38,025	463.7	84.1
							63	-6	-378	2,268	-166	996	27,556	437.4	83.9
							31	-7	-217	1,519	-88	616	7,744	249.8	83.7
							38	-8	-304	2,432	-126	1,008	15,876	417.8	83.2
2 5							11	-9	-99	891	-29	261	841	76.5	83.9
							5	-10	-50	500	-12	120	144	28.8	84.1
							24	-11	-264	2,904	-84	924	7,056	294.0	83.0
81	77	46	45	30	23	7	1,707	N	-792	28,616	363	17,300		11,894	
5	6	7	8	9	10	11			$\Sigma f_y d_y$	$\Sigma f_y d_y^2$	$\Sigma f_z d_z$	$\Sigma f_{xy} d_x d_y$			$\Sigma\left[\frac{(\Sigma' f_{xy} d_x)^2}{f_y}\right]$
405	462	322	360	270	230	77	363	$\Sigma f_x d_x$							
2,025	2,772	2,254	2,880	2,430	2,300	847	28,915	$\Sigma f_z d_z^2$							
178	211	177	207	153	148	52	-792	$\Sigma f_y d_y$							
890	1,266	1,239	1,656	1,377	1,480	572	17,300	$\Sigma f_{xy} d_x d_y$							
3,684	4,452	3,132.9	4,284.9	2,340.9	2,190.4	270.4									
391.2	578.2	681.1	952.2	780.3	952.3	386.3	11,194	$\Sigma\left[\frac{(\Sigma' f_{xy} d_x)^2}{f_x}\right]$							
3.40	3.58	3.95	4.20	4.37	4.81	5.14									

CORRECTIONS

$\frac{(\Sigma f_z d_z)(\Sigma f_y d_y)}{N} = -168$

$\frac{(\Sigma f_z d_z)^2}{N} = 77$

$\frac{(\Sigma f_y d_y)^2}{N} = 367$

$\log b = 4.4599651$ $\quad \log a = 4.2422432$ $\quad \log d = 4.0725072$ $\quad \log e = 4.0345081$

$\log c = 4.4510031$ $\quad \log \sqrt{\text{prod}} = 4.4554841$ $\quad \log b = 4.4599651$ $\quad \log c = 4.4510031$

$\log \text{prod} = 8.9109682$ $\quad \log r = 9.7867591$ $\quad \log \frac{d}{b} = 9.6125421$ $\quad \log \frac{e}{c} = 9.5835050$

$\log \sqrt{\text{prod}} = 4.4554841$ $\quad \frac{1}{2}\log \frac{d}{b} = 9.8062710$ $\quad \frac{1}{2}\log \frac{e}{c} = 9.7917525$

$r_{zy} = .6120$ $\quad \eta_{zy} = .6401$ $\quad \eta_{yz} = .6191$

4. Tests for Linearity

In order to determine whether or not a correlation table is sufficiently linear so that r may replace η as a measure of association, a test for linearity known as Blakeman's test may be applied. The observed difference between η_{yx} and r for the data on page 182 is $.6191 - .6120 = .0071$. With a similar table it appears quite likely that this difference might be zero.

The tests of Blakeman which we shall use here may be expressed in the following form: The difference between η and r is to be regarded as insignificant, provided that

$$\eta^2 - r^2 < \frac{4.047}{\sqrt{N}} \sqrt{(\eta^2 - r^2)\{(1 - \eta^2)^2 - (1 - r^2)^2 + 1\}}, \quad (67)$$

{Blakeman's test for linearity}

or, if $\eta^2 - r^2$ is small in comparison with r,

$$\sqrt{N}\sqrt{\eta^2 - r^2} < 4.047. \quad (68)$$

{Blakeman's short test for linearity}

A full discussion of such sampling tests will be found in Chapter XIII, but for the present we shall merely illustrate the above rules by applying them to the university and high-school correlations found in the preceding section.

For the coefficients η_{yx} and r we have, upon substituting the necessary values in formula (67),

$$.00874 < \frac{4.047}{41.32} \sqrt{.00874\{.989\}} = .00911.$$

Using formula (68), we have

$$41.32 \times .0935 = 3.86 < 4.047.$$

By both tests, therefore, the regression is to be regarded as linear and the use of a linear equation for predicting university from high-school grades is justified.

Applying formula (68) to η_{xy} and r, we find

$$41.32 \times .1876 = 7.75 > 4.047.$$

The regression in this case is non-linear, and for a full measure of the association, η_{xy} must be used.

For a small number of cases (say 50) it is frequently impossible to determine with certainty whether or not the regression is linear. With small tables, therefore, unless the regression is obviously curved, the calculation of r will be all that is required. With considerably larger bodies of data, however, the test becomes important and the use of r should be justified by comparison with both of the correlation ratios.

5. A Method of Eliminating the Effect of a Variable upon the Association between Two Others

If three or more correlated variables are involved, the association between two of them for a fixed value of the third is

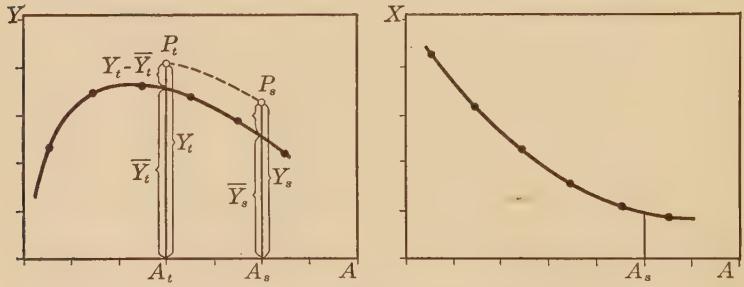


Fig. 45. Illustrating formula (69)

often required. In case the regressions are all linear throughout, the problem may be solved by the use of multiple correlation as shown in Chapter XV, but with non-linear relationships the solution becomes more difficult.

The most direct and the best method for dealing with such problems is to correct the two associated variables for values of the third. The method may be illustrated for two variables, both having non-linear correlation with age. Designating these as X, Y, and A, the correlation r_{xy} for A eliminated is required.

It is first necessary to prepare the correlation tables for X with A and Y with A and determine the regression curves for

Y on A and X on A as shown in Fig. 45. These curves may be drawn in free-hand, or fitted by the method of least squares as shown in Chapter XVI. A certain age, A_s, is then selected for both tables, and all the values of X and of Y are corrected to this age.

Let the ordinate of any observation in the table at age A_t be denoted by Y_t, and let $\overline{Y}_t$ and $\overline{Y}_s$ be the mean values of Y at A_t and A_s furnished by the regression curve. Then the required value of Y_s at A_s will be given by the relation

$$Y_s = \overline{Y}_s + (Y_t - \overline{Y}_t). \left\{ \begin{matrix} \text{Corrective formula} \\ \text{for eliminating age} \end{matrix} \right\} \quad (69)$$

From Fig. 45 it will be noted that this formula merely assumes that the growth in Y_t from A_t to A_s is parallel to the regression curve between these points. The corrected variable Y_s is thus the most likely value that Y_t will have when the individual has reached the standard age.

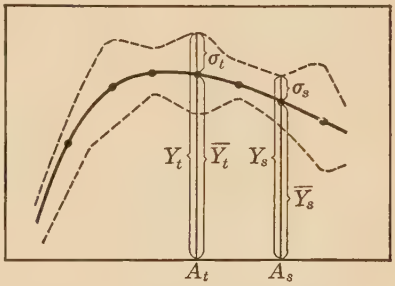


FIG. 46. Illustrating formula (70)

The arithmetic is most easily done by preparing a table of values $\overline{Y}_s$ for all ages and then applying formula (69) to the observations at each age across the correlation table. Similar corrections may be made for the variable X, and all the results recorded on the tabulation cards. The correlation between the corrected values Y_s and X_s then gives a good approximation to the result that would have been obtained if all the subjects had been measured at the same age, A_s.

In case the standard deviations of the arrays of ages are not equal, another correction may be made. Equal variability of the arrays across the table is described as *homoscedasticity*, and unequal variability as *heteroscedasticity*. The new correction, then, is for heteroscedasticity as illustrated in Fig. 46.

If an individual at A_t is one standard deviation above the mean $\bar{Y}_t$, his most probable deviation at age A_s will be one standard deviation above the mean at standard age. Denoting the standard deviations of the arrays at A_t and A_s by σ_t and σ_s, respectively, the corrective formula then becomes

$$Y_s = \bar{Y}_s + (Y_t - \bar{Y}_t)\frac{\sigma_s}{\sigma_t}.^{*} \quad \left\{\begin{array}{l}\text{Corrective formula adjusting} \\ \text{for age and heteroscedasticity}\end{array}\right\} \quad (70)$$

In applying this formula it is necessary to work out the ratios $\frac{\sigma_s}{\sigma_t}$ for each age, multiply the result by the corrective factor $(Y_t - \bar{Y}_t)$, and add to the value for $\bar{Y}_s$.

The corrected values X_s and Y_s may now be correlated, and the result will give the relationship between these variables for the age eliminated. This is essentially what is known as a *partial correlation* between X and Y (for A fixed). In case the variables X and Y have *linear* regression with A, a partial correlation may be worked out by the use of a formula (see Chapter XV).

It may finally be noted that the regression curves for the corrected variables X_s and Y_s may be non-linear and the correlation ratio required. Whatever measure of relationship is used, however, the resulting association is freed from the effect of the third variable A.

EXERCISES

1. Work out the correlation coefficient and the two correlation ratios for the table on page 187. Apply the tests for linearity. ($r = -.828$; $\eta_{xy} = .961$; $\eta_{yx} = .958$. *Ans.*)

2 and 3. Calculate the correlation coefficient and ratios for the tables on pages 188 and 189, and test for linearity.

4. Show that the method for correction given by formula (70) is equivalent to equating standard scores at ages A_t and A_s.

* For an illustration of the use of this formula see a paper by the author, "On the Relation of Vital Capacity to Certain Psychical Characters," *Biometrika*, Vol. XVI, p. 139.

Per Cent of Total School Revenue Derived from State Sources	Cost per Pupil in Dollars																				Total
	$ 4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	
100	1																				1
95																					—
90		1																			1
85																					—
80																					—
75		2																			2
70			1																		1
65			5																		5
60			6																		6
55			6	3																	9
50			3	6	2																11
45			1	13	3																17
40				6	13	5	1														25
35					7	13		1													21
30					1	4	13	4	2												24
25							5	6	1	4	1										17
20										5	3	2									10
15												1			3	1			1		6
10																	1				1
5																					
Total	1	3	22	28	26	22	19	11	3	9	4	3	—	—	3	1	1	—	1	157	

VARIABLES: ALPHA RAW SCORE × MENTAL AGE STANFORD-BINET EXAMINATION. GROUP X: SPECIAL EXPERIMENTAL GROUP *

ALPHA RAW SCORE	STANFORD-BINET MENTAL AGE †															
	5–5.9	6–6.9	7–7.9	8–8.9	9–9.9	10–10.9	11–11.9	12–12.9	13–13.9	14–14.9	15–15.9	16–16.9	17–17.9	18–18.9	19–19.9	f
180–189	..	..	..	..	..	..	..	..	..	..	..	..	1	1	1	3
170–179	..	..	..	..	..	..	..	..	..	..	..	..	..	1	..	1
160–169	..	..	..	..	..	..	..	..	..	..	..	1	1	4	1	7
150–159	..	..	..	..	..	..	..	..	..	..	1	1	2	1	..	5
140–149	..	..	..	..	..	..	..	..	..	..	1	2	4	4	1	12
130–139	..	..	..	..	..	..	..	..	..	1	2	..	5	9	..	17
120–129	..	..	..	..	..	..	..	..	1	3	1	9	4	2	1	21
110–119	..	..	..	..	..	..	..	..	1	4	3	6	6	1	..	21
100–109	..	..	..	..	..	..	..	..	2	1	6	7	3	3	..	22
90–99	..	..	..	..	..	..	..	3	6	5	8	8	5	4	..	39
80–89	..	..	..	..	..	1	2	7	6	11	12	5	7	1	1	53
70–79	..	..	..	..	..	1	3	6	4	11	7	4	2	..	..	38
60–69	..	..	..	..	..	..	3	7	11	9	11	6	4	2	..	53
50–59	..	..	..	..	2	1	8	7	4	10	4	..	2	..	..	38
40–49	..	..	..	..	2	5	13	12	8	8	2	2	1	..	..	53
30–39	..	..	..	2	3	9	15	15	11	9	2	1	..	1	..	68
20–29	..	..	..	..	11	13	9	10	10	4	2	1	..	..	..	60
10–19	..	..	..	5	15	18	8	9	4	..	1	..	..	..	..	60
0–9	1	2	1	15	29	18	8	5	1	1	..	1	..	..	..	82
f	1	2	1	22	62	66	69	81	69	77	63	54	47	34	5	653

$r = .793$;
$\eta_{xy} = .807$;
$\eta_{yx} = .799$.
Ans.

* Data from Memoirs of the National Academy of Sciences, Vol. XV, p. 621.

† In the original tables, from which computations were made, intervals of half-year mental ages were used. In the present table the intervals are a year, running from 5–5.9, 6–6.9, etc.

VARIABLES: BETA TOTAL SCORE × ALPHA TOTAL SCORE. GROUP X: SPECIAL EXPERIMENTAL GROUP*

BETA TOTAL \ ALPHA TOTAL	0–4	5–9	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99	100–104	105–109	110–114	115–119	120–124	125–129	130–134	135–139	140–144	145–149	150–154	155–159	160–164	165–169	170–174	175–179	180–184	185–189	f
115–119	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	2
110–114	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	1	.	1	.	1	.	.	.	.	.	.	7
105–109	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	3	.	1	.	1	.	.	.	.	.	.	9
100–104	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	1	3	2	3	1	1	1	.	1	2	1	4	1	.	1	2	1	2	1	2	.	.	.	.	30
95–99	.	.	.	.	.	.	.	.	.	1	.	1	.	2	2	1	3	1	3	2	1	2	3	2	1	3	2	1	3	2	2	1	1	2	.	1	.	.	42
90–94	.	.	1	.	.	.	.	1	.	.	1	.	2	4	4	4	.	3	1	1	2	1	3	2	6	1	4	2	1	3	2	2	.	1	1	.	.	.	58
85–89	.	.	.	.	.	.	.	.	1	2	.	4	7	3	5	3	6	4	4	8	6	2	.	4	4	2	1	1	3	2	2	.	2	1	1	.	1	.	80
80–84	.	.	.	.	.	.	2	.	5	4	4	3	3	11	6	7	9	3	7	2	4	3	3	3	1	6	.	2	2	2	.	2	.	.	.	.	.	.	93
75–79	.	.	1	.	1	.	5	4	4	4	4	4	7	6	2	2	10	5	2	4	5	1	4	3	3	.	3	2	2	1	1	.	.	.	.	.	.	.	80
70–74	.	.	.	1	1	3	6	6	5	6	5	5	6	5	7	4	5	4	3	4	5	3	1	1	1	1	.	1	.	1	.	.	.	.	.	.	.	.	76
65–69	.	1	1	2	1	2	3	7	5	4	3	5	8	3	2	4	1	4	4	3	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	71
60–64	.	.	3	2	3	6	7	6	9	9	6	5	2	4	4	1	4	.	.	.	.	3	.	.	1	2	1	.	.	.	.	.	.	.	.	.	.	.	74
55–59	.	2	1	2	8	8	8	4	6	3	3	1	7	1	4	.	.	.	.	.	.	.	1	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	62
50–54	.	.	1	5	7	8	8	6	2	3	5	6	2	.	2	.	3	.	1	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	59
45–49	2	3	5	4	5	6	3	4	2	2	.	2	.	1	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	41
40–44	3	6	6	2	9	6	4	6	1	1	5	1	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	51
35–39	2	6	8	5	7	5	5	4	2	1	1	2	1	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	51
30–34	5	3	4	2	5	2	2	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	26
25–29	13	6	1	4	3	3	1	2	4	.	.	1	.	.	2	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	41
20–24	14	9	7	2	3	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	37
15–19	10	4	2	.	2	.	.	1	.	1	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	21
10–14	8	4	3	1	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	18
5–9	5	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8
0–4	8	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10
f	70	48	44	33	55	50	56	55	46	42	37	40	46	40	41	26	43	31	25	41	20	22	13	20	14	21	10	13	9	7	6	5	4	7	2	1	1	3	1047

$r = .778$; $\eta_{xy} = .803$; $\eta_{yz} = .827$. *Ans.*

* Data from Memoirs of the National Academy of Sciences, Vol. XV, p. 621.

CHAPTER XI

THE BINOMIAL DISTRIBUTION

1. Introductory

As pointed out in the first chapter, the inductive side of statistical method is based on the theory of probability. The comparison of results from different samples, inferences regarding differences, and generalizations of various sorts are possible only by resorting to the theory of chance.

So important is this aspect of statistical science that some writers * devote practically all of their treatment to the theory of probability. For an elementary course and for the non-mathematical student such extensive treatment is impossible. We shall therefore be content to present here some of the simplest ideas in this theory with the understanding that the student is urged to amplify his knowledge of probability by consulting such works as Keynes,† Whittaker,‡ and Fisher.

In the present chapter we shall take up certain elementary theorems in probability and discuss the chance distribution known as the point binomial. Certain properties of this series which are important in the theory of sampling will also be considered. The binomial law also serves as a good introduction for the normal probability curve, which will be taken up in the following chapter.

In order to remind the student of some of the algebra useful in the development of the point binomial we shall turn first to the theory of combinations.

* Arne Fisher, The Mathematical Theory of Probabilities. The Macmillan Company, second edition, 1923.

† J. M. Keynes, A Treatise on Probability. The Macmillan Company, 1921.

‡ Whittaker and Robinson, The Calculus of Observations. D. Van Nostrand Company, 1924.

2. Permutations and Combinations

Suppose that a group of n objects is given. Any set of r* of these objects, without regard to their order, is called a *combination* of the n objects taken r at a time, and is denoted by the symbol ${}_nC_r$. For example, the combinations of the first four letters of the alphabet taken three at a time are

$$abc \qquad abd \qquad acd \qquad bcd$$

Since there are four of these, we may write ${}_4C_3 = 4$.

If the order of the objects be taken into account, the arrangements are known as *permutations* and are denoted by ${}_nP_r$. Thus the letters a, b, and c may be arranged in a row in the order abc, acb, bac, bca, cab, and cba, so that ${}_3P_3 = 6$. In the case of four letters, each of the four combinations of three furnishes six permutations, so that the total number of permutations of four things taken three at a time is twenty-four, or ${}_4P_3 = 24$.

The general formulas for permutations and combinations may be shown to have the forms

$$ {}_nP_r = n(n-1)(n-2)\cdots(n-r+1) \tag{71}$$

and

$$ {}_nC_r = \frac{n(n-1)(n-2)\cdots(n-r+1)}{1\cdot 2\cdot 3\cdots r} = \frac{{}_nP_r}{r!}. \tag{72}$$

The quantity $r!$ is known as "factorial r" and means the product of all integers from 1 to r.

It is also shown in algebra that ${}_nC_r = {}_nC_{n-r}$, so that ${}_nC_n = {}_nC_0 = 1$. This theorem will be needed in a later section.

Applying the above formulas to four letters taken two at a time, we find

$${}_4P_2 = 4 \times 3 = 12, \quad \text{and} \quad {}_4C_2 = \frac{4\times 3}{1\times 2} = 6.$$

* This "r" should not be confused with the correlation coefficient. It seemed best to retain this symbol in the theory of combinations because of its wide use by mathematicians.

These results may be easily verified by making all possible arrangements of four letters two at a time,

ab	*ac*	*ad*	*bc*	*bd*	*cd*
ba	*ca*	*da*	*cb*	*db*	*dc*

The student may also check the following numerical results by applying the above formulas:

$$_5P_3 = 60 \qquad _5C_3 = 10 \qquad _6P_4 = 360 \qquad _6C_4 = 15$$
$$_6P_2 = 30 \qquad _6C_2 = 15 \qquad _{10}P_3 = 720 \qquad _{10}C_3 = 120$$

3. Elementary Probability

If an event may happen in h ways and fail in k ways, and if each of the $h + k$ ways is equally likely to occur, the mathematical probability* of the event happening is

$$p = \frac{h}{h+k}, \tag{73}$$

and the probability of its failing is

$$q = \frac{k}{h+k}. \tag{74}$$

It is evident that the probability of an event happening plus the probability of its failing is equal to 1, which is the mathematical symbol for certainty. The above results may also be expressed by saying that the *odds* are h to k in favor of the event happening, or k to h against its occurrence.

Some of the simplest examples of such probability are furnished by the results of penny and dice tossing. Let us assume that the penny is a homogeneous disk and exclude the possibility of its standing upon an edge or sticking in a crack. If the turning up of the head is regarded as a successful event and the turning up of the tail as a failure, it is evident that $p = q = \frac{1}{2}$. In the case of the die, the turning up of the ace might be considered

* We are not concerned here with the various types of probability discussed in such treatises as Keynes, op. cit.

a success. Since there are six equally likely ways for the die to fall, the probability of the successful occurrence of this event is $\frac{1}{6}$, and that of its failure is $\frac{5}{6}$. It should also be noted that the odds are even of turning up a head with a coin, but are five to one against the turning up of an ace on a die.

If several pennies or dice are used, the resulting tosses are considered as *compound events*, and the occurrences of the individual events are regarded as entirely *independent* of one another. Thus with two pennies, a toss resulting in two heads is a compound event, and the fall of one penny is not influenced in any way by the fall of the other.

It should be observed that the same results will be obtained whether we deal with the occurrences of a number of similar events, or with several trials of the same event. This may be illustrated in the case of penny-tossing. The various tosses which occur with three coins are the same combinations that arise when one penny is tossed three times in succession and the individual occurrences then combined. A compound event may therefore be obtained by several trials of a single event.

The probability for the occurrence of a compound event such as all heads on three successive trials with a penny (or from one toss of three pennies) may be obtained by applying the definition of probability given above. The number of equally likely ways in which the coin may fall on the first trial is 2, and on each of the other two trials also 2, so that the total number of equally likely possible ways for the compound event to occur is $2 \times 2 \times 2 = 8$. The number of favorable ways for the event to happen is clearly 1, so that the required probability is $\frac{1}{8}$. By similar reasoning it may be shown that *if the probability of an event is* p, *the probability of its occurrence on all of* n *trials is* p^n. In case we are dealing with a number of dissimilar independent events whose individual probabilities are $p_1, p_2, p_3 \cdots$, the probability of their all occurring together is $p_1 \times p_2 \times p_3 \cdots$.

This last theorem may be illustrated in the case of a penny, a die, and a deck of playing cards. The probability of turning

up a head on the penny, an ace on the die, and the king of spades from the deck on one trial for each is $\frac{1}{2} \times \frac{1}{6} \times \frac{1}{52} = \frac{1}{624}$.

The probabilities of a complete set of compound events may be illustrated by examining the combinations which occur when a coin is tossed three times. If the coins are designated by 1, 2, and 3, while H stands for head and T for tail, the following arrangement of the eight different throws may be made:

(1)	T		T	T	H		T	H	H		H
(2)	T		T	H	T		H	T	H		H
(3)	T		H	T	T		H	H	T		H

Of the 8 equally likely combinations, one is TTT, or all tails, while another is TTH, or two tails and a head. This latter compound event may occur in 3 different ways, however, so that the probability of its occurrence is $\frac{3}{8}$. A complete set of such probabilities may then be set down as follows:

$$\text{Probability of } TTT = \tfrac{1}{8}$$
$$\text{Probability of } TTH = \tfrac{3}{8}$$
$$\text{Probability of } THH = \tfrac{3}{8}$$
$$\text{Probability of } HHH = \tfrac{1}{8}$$

A general expression for the above results may now be obtained by using the theorem for the probability of compound events. The probability that an event will occur on all of n trials is evidently p^n. In the above problem this is $(\frac{1}{2})^3 = \frac{1}{8}$. The probability that the event will occur $n-1$ times and fail once is $p^{n-1}q$. This result, however, may occur in n different ways, as is evident from the illustrative problem. The complete probability for $n-1$ successes and one failure is therefore $np^{n-1}q$. Next, the probability that in n trials the event will occur $n-2$ times and fail twice is $p^{n-2}q^2$. But again, this may occur in the number of ways in which two things may be selected from n, which is $\frac{n(n-1)}{1 \cdot 2} = {}_nC_2$. The total probability is therefore ${}_nC_2(p^{n-2}q^2)$. Thus for three trials the probability that

there will be one success and two failures is $\frac{3 \cdot 2}{1 \cdot 2} \times \frac{1}{2}\left(\frac{1}{2}\right)^2 = \frac{3}{8}$ in this example.

Continuing in the same way, it is evident that the general expression for the probability of obtaining exactly r successes and $(n - r)$ failures is given by ${}_nC_r p^r q^{n-r}$.

4. The Binomial Theorem and the Point Binomial

The binomial theorem may be written in the form

$$(a + b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{1 \cdot 2} a^{n-2}b^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} a^{n-3}b^3 + \cdots + b^n. \quad (75)$$

The expansion on the right is the general result of multiplying out $(a + b)(a + b)(a + b)$ to n factors. By making use of the notation for combinations, a more convenient form of this expansion may be obtained:

$$(a+b)^n = {}_nC_0 a^n + {}_nC_1 a^{n-1}b + {}_nC_2 a^{n-2}b^2 + {}_nC_3 a^{n-3}b^3 + \cdots + {}_nC_n b^n. \quad (76)$$

Applying this theorem to $(q + p)^n$, we have

$$(q+p)^n = {}_nC_0 q^n + {}_nC_1 q^{n-1}p + {}_nC_2 q^{n-2}p^2 + {}_nC_3 q^{n-3}p^3 + \cdots + {}_nC_n p^n, \quad (77)$$

{Point binomial}

the terms of which agree with the general expression for the probability of r successes found in the preceding section. The conclusion then is that *if* n *trials be made of an event for which the probability of occurrence is* p *and the probability of failure is* q, *the probabilities of* $0, 1, 2, \cdots n$ *successes are given by the successive terms in the expansion of the binomial* $(q + p)^n$.

As an illustration of this theorem the thirteen terms of the binomial $(\frac{1}{2} + \frac{1}{2})^{12}$ are worked out in Table 36 on page 196. These are the probabilities of getting 0, 1, 2, $\cdots$ 12 heads when one coin is tossed twelve times or twelve coins are tossed once.

It is apparent from these results that the probability of getting all heads or all tails is very small. If twelve coins were used, only about once in 4000 throws would such an event occur.

TABLE 36. ILLUSTRATING THE PROBABILITIES OF OBTAINING 0, 1, 2 . . . HEADS IN TOSSING TWELVE COINS

SUCCESSES (HEADS)	PROBABILITIES
0	$\frac{1}{4096} = .000244$
1	$\frac{12}{4096} = .002930$
2	$\frac{66}{4096} = .016113$
3	$\frac{220}{4096} = .053711$
4	$\frac{495}{4096} = .120850$
5	$\frac{792}{4096} = .193359$
6	$\frac{924}{4096} = .225586$
7	$\frac{792}{4096} = .193359$
8	$\frac{495}{4096} = .120850$
9	$\frac{220}{4096} = .053711$
10	$\frac{66}{4096} = .016113$
11	$\frac{12}{4096} = .002930$
12	$\frac{1}{4096} = .000244$
Total	1 1.000000

The expression $(q + p)^n$ is often called the *point binomial,* since its expansion is represented by a series of isolated points. In Fig. 47 these points have been connected by straight lines, forming a polygon very much like the normal curve in general appearance (see Chapter XII).

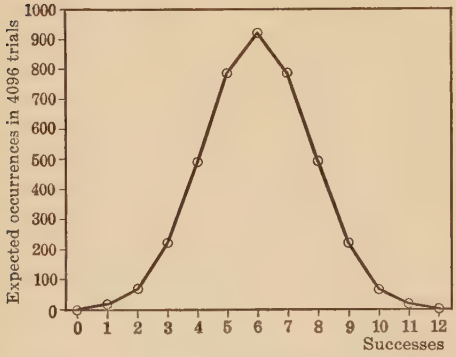


FIG. 47. Plot of the binomial $(\frac{1}{2} + \frac{1}{2})^{12}$

It has already been proved that the probability of a specified number of successes is given by the appropriate term in the point binomial. Another important result is that *the probability of an event occurring* **r** *or more times in* **n** *trials is the sum of the terms in the expansion of* $(q + p)^n$ *from* ${}_nC_rq^{n-r}p^r$ *to* ${}_nC_np^n$ *inclusive.* This follows from the fact that the $n + 1$ compound events are mutually exclusive, or such that the occurrence of one com-

bination excludes, for that throw, the other n possible arrangements. As shown in algebra, the probability that some one or other of such mutually exclusive events will occur is the sum of the probabilities of the separate (here compound) events.

Thus if twelve coins are thrown the probability of obtaining nine or more heads on a single toss is the sum of the probabilities $\frac{220}{4096} + \frac{66}{4096} + \frac{12}{4096} + \frac{1}{4096} = \frac{299}{4096}$, or .073. This result may also be worked out by noting that of the 4096 equally likely arrangements of the 12 coins there are 220 ways in which nine heads may turn up, 66 ways in which ten heads may occur, 12 ways for eleven heads to appear, and 1 way in which twelve heads may be obtained. This gives a total of 299 ways in which at least nine heads may appear, and the probability for such an occurrence is $\frac{299}{4096}$ from the definition of simple probability.

5. The Mean of the Point Binomial and its Standard Deviation

We shall next prove two interesting theorems in connection with the point binomial. These are known as the theorems of Bernoulli and are of great importance in statistical theory.

The mean of the point binomial is np, and its standard deviation is $\sqrt{npq}$.

In proving the theorems, M and σ are calculated as follows:

Table 37. Calculation of M and σ for the Point Binomial

Successes	Frequency	d	fd	fd^2
0	q^n	0	–	–
1	$nq^{n-1}p$	1	$nq^{n-1}p$	$nq^{n-1}p$
2	$\frac{n(n-1)}{1\cdot 2}q^{n-2}p^2$	2	$n(n-1)q^{n-2}p^2$	$2\,n(n-1)q^{n-2}p^2$
3	$\frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}q^{n-3}p^3$	3	$\frac{n(n-1)(n-2)}{1\cdot 2}q^{n-3}p^3$	$3\,\frac{n(n-1)(n-2)}{1\cdot 2}q^{n-3}p^3$
–	–	–	–	–
–	–	–	–	–
–	–	–	–	–
–	–	–	–	–
Totals . .	1		np	$np[1 + p(n-1)]$

The sum of the frequencies is $(q+p)^n$, or unity, and Σfd may be readily factored into

$$np\left[q^{n-1}+(n-1)q^{n-2}p+\frac{(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2+\cdots\right]$$
$$=np(q+p)^{n-1}=np.$$

We may now apply the formula for the mean, $M=A+\left(\frac{\Sigma fd}{N}\right)h$. Since $A=0$, $N=1$, $\Sigma fd=np$, and $h=1$, the mean of the binomial becomes

$$M=np. \quad \left\{\begin{matrix}\text{Mean of the point}\\ \text{binomial}\end{matrix}\right\} \qquad (78)$$

In order to obtain the standard deviation it is necessary to find Σfd^2 for the above series. The last column of items in Table 37 may be factored as follows:

$$\Sigma fd^2=np\left[q^{n-1}+2(n-1)q^{n-2}p+\frac{3(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2+\cdots\right].$$

The terms in the brackets may now be broken up to form two series in $(q+p)$. Thus,

$$\Sigma fd^2=np\left[\left\{q^{n-1}+(n-1)q^{n-2}p+\frac{(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2+\cdots\right\}\right.$$
$$\left.+\left\{(n-1)q^{n-2}p+2\,\frac{(n-1)(n-2)}{1\cdot 2}q^{n-3}p^2+\cdots\right\}\right]$$
$$=np[(q+p)^{n-1}+(n-1)p\,\{q^{n-2}+(n-2)q^{n-3}p+\cdots\}]$$
$$=np[(q+p)^{n-1}+(n-1)p(q+p)^{n-2}]$$
$$=np[1+(n-1)p].$$

Applying formula (17) for standard deviation, we find that

$$\sigma=\sqrt{\frac{np[1+(n-1)p]}{1}-n^2p^2}=\sqrt{np(1-p)},$$

or $$\sigma=\sqrt{npq}. \quad \{\text{Standard deviation of the point binomial}\} \qquad (79)$$

The above formulas make possible the complete description of certain distributions given by chance. The terms in the series furnish the ordinates of the curve, while the mean and the standard deviation from the formulas (78) and (79) are convenient measures of the central tendency and dispersion of such a distribution.

For the binomial $(\frac{1}{2} + \frac{1}{2})^{12}$ the mean and the standard deviation by these formulas work out at 6 and $\sqrt{3}$ respectively. In the case of such a symmetrical series, the mean is of course obtained by inspection.

If twelve dice are thrown and the turning up of an ace is considered a success, the probabilities of 0, 1, $2 \cdots 12$ successes are given by the terms in the expansion of $(\frac{5}{6} + \frac{1}{6})^{12}$. This series is distinctly skew, but the mean and the standard deviation are readily found to be 2 and $\sqrt{\frac{5}{3}}$ on applying formulas (78) and (79). Practical evidence of the convenience of these formulas may be obtained by working out the same results directly from the frequencies.

6. Experimental Verification of the Binomial Law

In order to see whether or not the actual results of penny and dice tossing come out as predicted by the above formulas, it will be interesting to cite one or two examples. While such experiments serve to verify in a rough way the properties of the point binomial, it should be noted that strictly speaking they are not verifications at all because the conditions implied in the formulas can never be met on actual trial. The perfectly homogeneous penny or die does not exist, nor is it possible to make the tosses so that certain throws are not favored over certain others. Differences between the observed trials and the theoretically correct results will then be due not only to the number of trials or size of the sample but to imperfections in the objects thrown, and to faulty methods in tossing them. The student is urged, however, to make a few personal experiments such as those quoted below in order that he may become more familiar with the meaning and practical utility of the binomial law.

In the following experiment twelve dice were thrown 4096 times, the method being to roll them down an inclined gutter of corrugated paper. A throw of 4, 5, or 6 was considered a

success, so that $p = q = \frac{1}{2}$. The theoretical mean will then be np, or 6, and the standard deviation $\sqrt{npq}$, or 1.732. The following table gives the observed and theoretical frequencies.

TABLE 38. OBSERVED AND THEORETICAL FREQUENCIES OF 0, 1, 2 ... SUCCESSES FROM THE TOSSING OF TWELVE DICE WITH THROWS OF FOUR, FIVE, OR SIX AS SUCCESSES

SUCCESSES	OBSERVED FREQUENCY	THEORETICAL FREQUENCY	SUCCESSES	OBSERVED FREQUENCY	THEORETICAL FREQUENCY
0	—	1	7 . . .	847	792
1	7	12	8 . . .	536	495
2	60	66	9 . . .	257	220
3	198	220	10 . . .	71	66
4	430	495	11 . . .	11	12
5	731	792	12 . . .	—	1
6	948	924	Total . .	4096	4096

The mean of the observed distribution is 6.139 and its standard deviation is 1.712. The actual proportion of successes is 0.512 instead of 0.5. The agreement, on the whole, is therefore rather good.

In the next experiment a throw of a 6 was considered a success, so that $p = \frac{1}{6}$, and $q = \frac{5}{6}$. The theoretical mean is 2 and the standard deviation is 1.291. The observed frequency distribution was as follows:

TABLE 39. OBSERVED FREQUENCIES OF 0, 1, 2 . . . SUCCESSES RESULTING FROM THE THROWS OF TWELVE DICE WITH THE TURNING OF A SIX AS A SUCCESS

SUCCESSES	FREQUENCY	SUCCESSES	FREQUENCY
0	447	5	115
1	1145	6	24
2	1181	7	7
3	796	8	1
4	380	Total.	4096

The observed mean is 2.000 and standard deviation 1.296, while the actual proportion of successes is .1667, agreeing with the theoretical values to an extent that is probably accidental.

The above results show that with careful extensive experiments such as these, the observed series is in good agreement with the binomial expansion.

7. The Binomial Applied to Statistical Data

In the case of frequency distributions of observed data affected by many factors, the point binomial might often be used were it not for the large number of terms involved, and the difficulty of replacing the mathematical probability, known *a priori*, by an empirical probability ratio furnished by the data.

As an illustration we may take the records of 400 candidates for the master's degree in a certain university. Among other requirements it was necessary for the candidate to have an average of *B*— or better. For the present purposes, such an average may be considered a success, and a lower average may be regarded as a failure. Out of 400 candidates 331 maintained a satisfactory average, so that the empirical probability of such a success is $\frac{3\,3\,1}{4\,0\,0} = .8275$. It should be noted that such a ratio might change considerably from time to time, and would also tend to be unstable when applied to small numbers. We cannot expect, therefore, to get as good results from such empirical ratios as from the probabilities in the case of penny-tossing.

The average number of candidates coming up at one time was about ten. Taking this number as the size of the sample (corresponding to the number of coins tossed) the point binomial $(.1725 + .8275)^{10}$ might be used to determine the probability for any number of successes, say nine or more.

The terms in this binomial (computed by logarithms) together with the results actually found by trial are given in the table on page 202. The probability of getting 9 or more successes in a sample of 10 is the sum of the probabilities .314 and .150, or .464. The expected number from 400 candidates will, therefore, be $400 \times .464$, or 186. This result happens to be in close agreement with the observed number, $(6 + 13)10 = 190$.

TABLE 40. OBSERVED AND THEORETICAL FREQUENCIES FOR THE NUMBER OF SUCCESSFUL CANDIDATES FOR THE MASTER'S DEGREE, IN SAMPLES OF TEN, THE TOTAL NUMBER OF CANDIDATES BEING 400

SUCCESSFUL CANDIDATES OUT OF TEN	OBSERVED FREQUENCY	THEORETICAL FREQUENCY	PROBABILITIES
10	6	6.0	.150
9	13	12.6	.314
8	12	11.8	.294
7	7	6.5	.164
6	0	2.4	.060
5	1	0.6	.015
4	1	0.1	.003
3	0	0.0	.000
Total	40	40	1.000

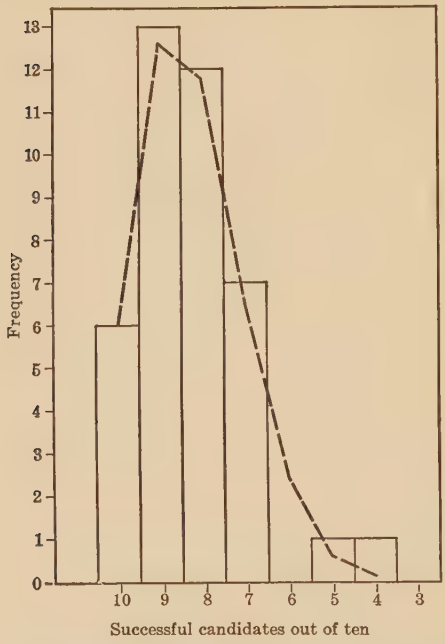


FIG. 48. Comparison of theoretical and observed frequencies for candidate data

The complete set of theoretical frequencies is found by multiplying the probability values by 40. These frequencies agree fairly well with those given by the data as shown in the above table and in Fig. 48.

Further evidence of the agreement of the two series may be found by comparing the theoretical and observed standard deviations. The former is $\sqrt{npq}$, or 1.19, while the latter is 1.28. The difference, or 0.09, may be readily accounted for by chance fluctuations in sampling (see formula (91) and the testing of differences in Chapter XIII).

EXERCISES

1. Expand the following binomials, and plot the results.

$$(\tfrac{1}{2}+\tfrac{1}{2})^7,\ (\tfrac{1}{2}+\tfrac{1}{2})^{10},\ (\tfrac{1}{3}+\tfrac{2}{3})^6,\ (.1+.9)^5,\ (.1+.9)^{10}.$$

2. If the terms in the expansion of $(\frac{1}{2}+\frac{1}{2})^{10}$ represent the probabilities of 0, 1, 2, 3 . . . 10 successes, find the probability of obtaining seven or more successes in ten trials. $(\frac{1}{1}\frac{7}{0}\frac{6}{2}\frac{}{4} = .172.$ *Ans.*)

3. Find the means and standard deviations for the binomials of Exercise 1, using formulas (78) and (79). Verify some of the answers by direct calculation from the full expansions of the binomials.

4. From Table 41 of Chapter XII determine the empirical probability of a man selected at random being over $71\frac{5}{16}$ inches in height. Use the total distribution. (.039. *Ans.*) What is the probability that a man's height will be between $66\frac{5}{16}$ and $67\frac{5}{16}$ inches? (.155. *Ans.*) What are the odds in favor of the latter occurrence? (Approx. 4 : 1. *Ans.*)

5. Suppose that a penny is tossed, a die thrown, and a card drawn from an ordinary deck. What is the probability of the combined event: head on the coin, ace or six on the die, and a heart on the card, with a single trial for each? $(\frac{1}{2}\frac{}{4}.$ *Ans.*)

6. What is the probability of turning up a total of eight with two dice? $(\frac{5}{3}\frac{}{6}.$ *Ans.*)

7. If three cards are drawn from a suit of thirteen cards, what is the chance that both king and queen are drawn? $(\frac{1}{2}\frac{}{6}.$ *Ans.*)

8. Show that if np be a whole number, the mean of the binomial coincides with the greatest term.

9. Derive formulas (78) and (79) by differentiating the expression $(q + px)^n$ with respect to x and setting $x = 1$.

CHAPTER XII

THE NORMAL PROBABILITY CURVE

1. Introductory

In the present chapter we shall discuss the properties and uses of the normal probability curve, the general form of which is doubtless already familiar to the student (see Fig. 51).

An example of a distribution resembling the normal probability curve is furnished by the mental age data in Fig. 49. When these data are separated into "normals" and "defectives" two fairly symmetrical curves result. Burt* explains the lack of complete symmetry in the curve for normals on the ground that the Binet Scale lacks adequate tests for the brighter children of the older ages. He concludes that even though his data are somewhat irregular, they do not "in any way contradict the hypothesis of 'normality,' the theory that ability is distributed in close conformity with the normal curve of error."

In the case of certain physical characteristics such as height, the normal curve appears to give an excellent fit to the observations. The data in Table 41, quoted from Yule,† furnish a very good example.

The histogram for the frequencies in the total column of the table is shown in Fig. 50, where the symmetry and general resemblance to the normal curve are apparent.

The above examples suggest that the frequency distributions of some mental and physical traits conform fairly well to the normal curve. It would be far from correct, however, to assume that all human characteristics are normally distributed. This assumption was made by an early statistician named Quetelet.

* Cyril Burt, Mental and Scholastic Tests, p. 162. King and Son, Ltd., London, 1921.

† Yule, Introduction to Statistics, p. 88.

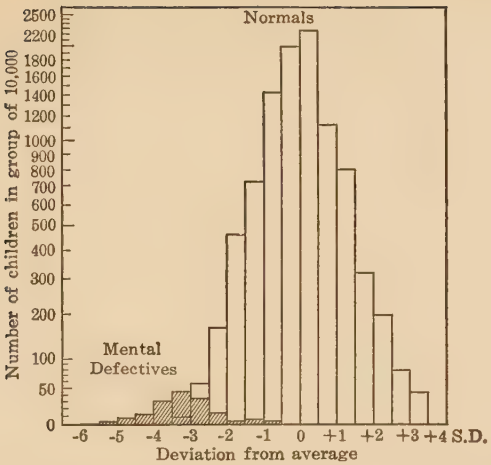


FIG. 49. Distribution according to general intelligence of children of ordinary elementary and special M.D. schools

From "Mental and Scholastic Tests," by Cyril Burt. Courtesy of P. S. King and Son, Ltd.

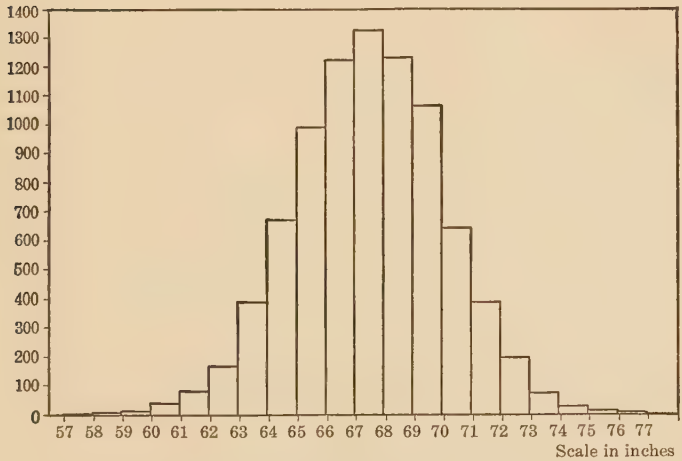


FIG. 50. Histogram for heights of 8585 men

TABLE 41. DISTRIBUTION OF STATURE FOR ADULT MALES BORN IN THE BRITISH ISLES

HEIGHT IN INCHES (WITHOUT SHOES)	NUMBER OF MEN ACCORDING TO BIRTHPLACE				TOTAL
	England	Scotland	Wales	Ireland	
$76\frac{15}{16}$–$77\frac{15}{16}$	1	1	—	—	2
$75\frac{15}{16}$–$76\frac{15}{16}$	1	4	—	—	5
$74\frac{15}{16}$–$75\frac{15}{16}$	9	6	1	—	16
$73\frac{15}{16}$–$74\frac{15}{16}$	16	15	1	—	32
$72\frac{15}{16}$–$73\frac{15}{16}$	48	26	2	3	79
$71\frac{15}{16}$–$72\frac{15}{16}$	117	69	6	10	202
$70\frac{15}{16}$–$71\frac{15}{16}$	254	102	21	15	392
$69\frac{15}{16}$–$70\frac{15}{16}$	473	115	33	25	646
$68\frac{15}{16}$–$69\frac{15}{16}$	753	218	52	40	1063
$67\frac{15}{16}$–$68\frac{15}{16}$	886	210	72	62	1230
$66\frac{15}{16}$–$67\frac{15}{16}$	918	210	128	73	1329
$65\frac{15}{16}$–$66\frac{15}{16}$	881	139	145	58	1223
$64\frac{15}{16}$–$65\frac{15}{16}$	740	109	108	33	990
$63\frac{15}{16}$–$64\frac{15}{16}$	524	47	83	15	669
$62\frac{15}{16}$–$63\frac{15}{16}$	320	19	48	7	394
$61\frac{15}{16}$–$62\frac{15}{16}$	128	9	30	2	169
$60\frac{15}{16}$–$61\frac{15}{16}$	70	2	9	2	83
$59\frac{15}{16}$–$60\frac{15}{16}$	39	2	—	—	41
$58\frac{15}{16}$–$59\frac{15}{16}$	12	—	1	1	14
$57\frac{15}{16}$–$58\frac{15}{16}$	3	1	—	—	4
$56\frac{15}{16}$–$57\frac{15}{16}$	1	—	1	—	2
Total	6194	1304	741	346	8585

He pictured an average man with physical and social traits at the means of a series of probability curves. The work of such men as Pearson and Charlier, however, has since shown that these characteristics are best represented by a variety of curves among which the probability curve is a special type. (See sections 8 and 9 of Chapter XVI.)

It will be shown in section 5 that the resemblance of a frequency distribution to the normal curve cannot be satisfactorily determined by mere inspection of the data. A rigorous test of the normality of a given distribution involves the superposition of a normal curve on the data and a mathematical comparison of the observed and theoretical frequency.

The normal probability curve is very important in the field of educational measurements because of its usefulness in scale construction and in many calculations involving qualitative series. It is usually necessary in such problems to assume some form of distribution and the normal curve is taken because, of all the curves which might be employed, it gives the best single approximation to the ordinary test score distribution. The mathematical properties of the probability curve, including tabulations of its integral and ordinate, make the calculations involved very much simpler than with some skew form of curve.

Although no formal derivation of the normal curve will be given, its relation to the point binomial will be shown as well as its usefulness in the elementary theory of probability.

2. The Equation of the Normal Probability Curve

As already pointed out, the practical use of the point binomial requires a great deal of labor. If, for example, the samples in the problem of section 7, Chapter XI, had consisted of twenty instead of ten candidates, the terms in the binomial $(q + p)^{20}$ would have to be computed.

An important simplification of the binomial law may now be reached by allowing the size of n to increase indefinitely. It is obvious from the binomials discussed thus far that as n becomes larger the resulting polygon over the $n + 1$ points becomes smoother and tends to spread out more and more in both directions from the mean. The limit to the point binomial, $(q + p)^n$, as n increases indefinitely, may be shown by mathematical proof * to be given by the continuous curve

$$y = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \quad \left\{\begin{matrix}\text{Normal curve}\dagger \\ \text{with area} = 1\end{matrix}\right\} \tag{80}$$

* Yule, Introduction to Statistics, p. 301 (simple proof).

† The normal probability curve was first given by De Moivre in 1733 but was later rediscovered by Laplace and Gauss.

where $e = 2.7183 \cdots$, which is the base of the Napierian system of logarithms, and π is the familiar ratio of the circumference of a circle to its diameter.

Just as the sum of the ordinates in the point binomial $(q + p)^n$ is equal to unity, so the area under this curve is equal to 1. From equation (80) it is evident that for $x = 0$, $y = \dfrac{1}{\sqrt{2\pi}\sigma}$; and that about the value 0, which is the mean, the curve is symmetrical, because the same positive and negative values of x give a single value for y. By writing the equation in the form

$$y = \frac{1}{\sqrt{2\pi}\sigma e^{+\frac{x^2}{2\sigma^2}}}, \quad (81)$$

it is also apparent that no matter how large or small x is taken, y will never become equal to zero. The curve is thus symmetrical about the mean at $x = 0$, and extends indefinitely in both directions, approaching the x-axis as an *asymptote* as shown in Fig. 51.

FIG. 51. Normal curve with unit area (if $\sigma = 1$)

In case the normal curve is applied to data for which the total frequency is N and not unity, the form of the equation becomes

$$y = \frac{N}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} = y_0 e^{-\frac{x^2}{2\sigma^2}}, \quad \left\{\begin{matrix}\text{Normal curve} \\ \text{with area} = N\end{matrix}\right\} \quad (82)$$

each of the ordinates for unit area being multiplied by N. The coefficient $\dfrac{N}{\sqrt{2\pi}\sigma}$ is often designated as y_0, and is the maximum ordinate at $x = 0$, since $e^0 = 1$ as noted in Chapter IV, section 4.

3. The Area, Ordinates, and Deviates of the Normal Curve

If the standard deviation of the normal curve be chosen as 1, for convenience, the equation then takes the form

$$z = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \quad \left\{ \begin{array}{l} \text{Ordinate of the normal curve, with} \\ \text{unit area and standard deviation} \end{array} \right\} \quad (83)$$

The values of x, or number of standard deviations from the mean, are called *deviates*; z is the usual symbol for the *ordinate* at a given deviate; and $\frac{1}{2}\alpha$ will be used to denote the *area* from the mean to such a deviate. These three functions of the curve have been computed and tabled in various ways, and are of the greatest importance for a variety of statistical calculations. An illustration of these functions is given in Fig. 52, the numbers being taken from Table 42. It will be noted that for a deviate $x = 1.5$, the ordinate z will have the value .130, while the area from the mean, or $\frac{1}{2}\alpha$, will be 43.3 per cent of the total area of the curve.

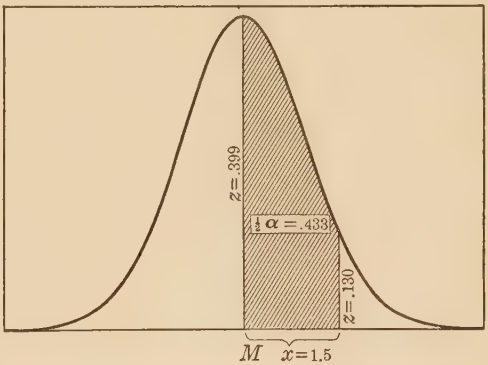


Fig. 52. Illustrating area, ordinates, and deviates for a normal curve

The methods for calculating the areas and deviates are a part of the calculus, but the ordinates may be determined by merely substituting various values for x in equation (83). For example, when $x = 0$, $z = \frac{1}{2.5066} = .3989$. Similarly, when $x = 1$, $z = \frac{1}{2.5066}(2.7183)^{-\frac{1}{2}} = .2420$.

For certain problems, which will be taken up later, it has been found convenient to calculate and table these functions in two ways:

(1) Areas and ordinates for given deviates, and

(2) Deviates and ordinates for given areas.

Complete tables for these values are found in Pearson's* "Tables for Statisticians and Biometricians," in Kelley's† "Statistical Method," and in more abbreviated form in a handbook prepared by the writer.‡ Two short lists of three-place values are also given in Tables 42 and 43.

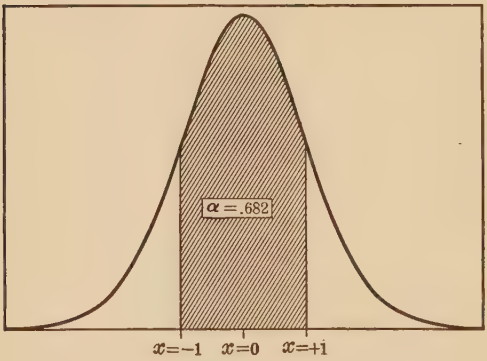


FIG. 53. Illustrating α for a normal curve

It is apparent that α is the area from $-x$ to $+x$, as shown in the accompanying figure. When $x = \pm 1$, $\alpha = .682$, from which it follows that more than two thirds of the total area under the curve is included between these limits. When $x = \pm 3$, $\alpha = .998$, showing that a range of 6σ includes more than 99 per cent of the frequency. It will also be noted that the ordinate at $x = 3$ is very small, being only $\frac{4}{399}$ of y_0, or .01 of the maximum ordinate at the mean.

Table 43 for deviates and ordinates in terms of area from the mean shows that for equal increments of $\frac{1}{2}\alpha$ there is very little change in x and z in the neighborhood of the mean, but very rapid change toward the extremities of the curve. For $\frac{1}{2}\alpha = .50$ the ordinate is equal to zero, and the deviate is infinite.

* Tables for Statisticians and Biometricians, edited by Karl Pearson. Cambridge University Press, England. Second edition, 1924.

† T. L. Kelley, Statistical Method. The Macmillan Company, 1923.

‡ Karl J. Holzinger, Statistical Tables for Students in Education and Psychology. The University of Chicago Press, 1925.

TABLE 42. AREAS AND ORDINATES FOR GIVEN DEVIATES FROM THE MEAN

x	$\frac{1}{2}\alpha$	z	x	$\frac{1}{2}\alpha$	z
0.0	.000	.399	2.1	.482	.044
0.1	.040	.397	2.2	.486	.035
0.2	.079	.391	2.3	.489	.028
0.3	.118	.381	2.4	.492	.022
0.4	.155	.368	2.5	.494	.018
0.5	.191	.352	2.6	.495	.014
0.6	.226	.333	2.7	.497	.010
0.7	.258	.312	2.8	.497	.008
0.8	.288	.290	2.9	.498	.006
0.9	.316	.266	3.0	.499	.004
1.0	.341	.242	3.1	.499	.003
1.1	.364	.218	3.2	.499	.002
1.2	.385	.194	3.3	.500	.002
1.3	.403	.171	3.4	.500	.001
1.4	.419	.150	3.5	.500	.001
1.5	.433	.130	3.6	.500	.001
1.6	.445	.111	3.7	.500	.000
1.7	.455	.094	3.8	.500	.000
1.8	.464	.079	3.9	.500	.000
1.9	.471	.066	4.0	.500	.000
2.0	.477	.054	4.1	.500	.000

When $\frac{1}{2}\alpha = .25$ it will be noted that $x = .674$, or, more exactly, $x = .6744898$. This value, which is known as the *probable error*, is therefore given by the relation

$$P.E. = .6744898\,\sigma. \quad \left\{\begin{matrix}\text{Relation between} \\ P.E. \text{ and } \sigma\end{matrix}\right\} \quad (84)$$

It is very frequently used as a unit of measurement on the normal scale instead of σ, chiefly because of long usage.

It may also be observed that $P.E.$ and Q are the same for a normal curve, since exactly half of the area is included when they are laid off on either side of the mean. With actual data, $P.E.$ will not be equal to Q, so that it is usually better to avoid the use of the term *probable error* in describing an observed frequency distribution. The term arose in connection with distributions of error such as those in astronomical measurements. With ordinary data such a deviate does not represent an error and the term *probable error* is therefore a misnomer.

TABLE 43. DEVIATES AND ORDINATES FOR GIVEN AREA FROM THE MEAN

$\frac{1}{2}\alpha$	x	z	$\frac{1}{2}\alpha$	x	z
.00	0.000	.399	.26	0.706	.311
.01	0.025	.399	.27	0.739	.304
.02	0.050	.398	.28	0.772	.296
.03	0.075	.398	.29	0.806	.288
.04	0.100	.397	.30	0.842	.280
.05	0.126	.396	.31	0.878	.271
.06	0.151	.394	.32	0.915	.262
.07	0.176	.393	.33	0.954	.253
.08	0.202	.391	.34	0.994	.243
.09	0.228	.389	.35	1.036	.233
.10	0.253	.386	.36	1.080	.223
.11	0.279	.384	.37	1.126	.212
.12	0.305	.381	.38	1.175	.200
.13	0.332	.378	.39	1.227	.188
.14	0.358	.374	.40	1.282	.175
.15	0.385	.370	.41	1.341	.162
.16	0.412	.366	.42	1.405	.149
.17	0.440	.362	.43	1.476	.134
.18	0.468	.358	.44	1.555	.119
.19	0.496	.353	.45	1.645	.103
.20	0.524	.348	.46	1.751	.086
.21	0.553	.342	.47	1.881	.068
.22	0.583	.337	.48	2.054	.048
.23	0.613	.331	.49	2.326	.027
.24	0.643	.324	.50	∞	.000
.25	0.674	.318			

4. COMPARISON OF THE POINT BINOMIAL AND THE NORMAL CURVE

The close agreement between the binomial series and the normal curve may be illustrated for the binomial $(\frac{1}{2}+\frac{1}{2})^{16}$, the ordinates for which are given by expansion as shown in Chapter XI.

In order to compute the normal ordinates at the 17 binomial points it is first necessary to calculate the values of the latter as deviates from the mean. Since the standard deviation of the binomial is $\sqrt{npq}$, or 2, for the above series, the deviates at 0, 1, 2, 3, $\cdots$ successes will be $\frac{0-8}{2}=-4$, $\frac{1-8}{2}=-3.5$, $\frac{2-8}{2}=-3$, etc.

The ordinates of the normal curve for these deviates may now be looked up in Table 42, and divided by 2 in order to make them comparable with the binomial ordinates. The tabled values, of course, are for unit standard deviation. A complete list of the abscissas and ordinates for both curves may then be obtained as shown in Table 44.

TABLE 44. ORDINATES FOR THE BINOMIAL $(\frac{1}{2} + \frac{1}{2})^{16}$, WITH CORRESPONDING NORMAL ORDINATES

SUCCESSES	BINOMIAL ORDINATES	$\frac{x}{\sigma}$	NORMAL ORDINATES FOR $\sigma = 1$	NORMAL ORDINATES FOR $\sigma = 2$
0	.000	− 4.0	.000	.000
1	.000	− 3.5	.001	.0005
2	.002	− 3.0	.004	.002
3	.0085	− 2.5	.018	.009
4	.028	− 2.0	.054	.027
5	.067	− 1.5	.130	.065
6	.122	− 1.0	.242	.121
7	.1745	− 0.5	.352	.176
8	.196	0.0	.399	.199
9	.1745	+ 0.5	.352	.176
10	.122	+ 1.0	.242	.121
11	.067	+ 1.5	.130	.065
12	.028	+ 2.0	.054	.027
13	.0085	+ 2.5	.018	.009
14	.002	+ 3.0	.004	.002
15	.000	+ 3.5	.001	.0005
16	.000	+ 4.0	.000	.000
Total	1.000			1.000

From these values and by inspection of Fig. 54 it is apparent that the agreement between the two curves is very close. For more terms, of course, the discrepancies between the ordinates would have been even less than those found here.

The equation of the normal curve here considered is clearly

$$y = \frac{1}{\sqrt{8\pi}} e^{-\frac{x^2}{8}},$$

since it is only necessary to substitute $\sigma = 2$ in equation (80). The mean of the curve is set at 8 successes.

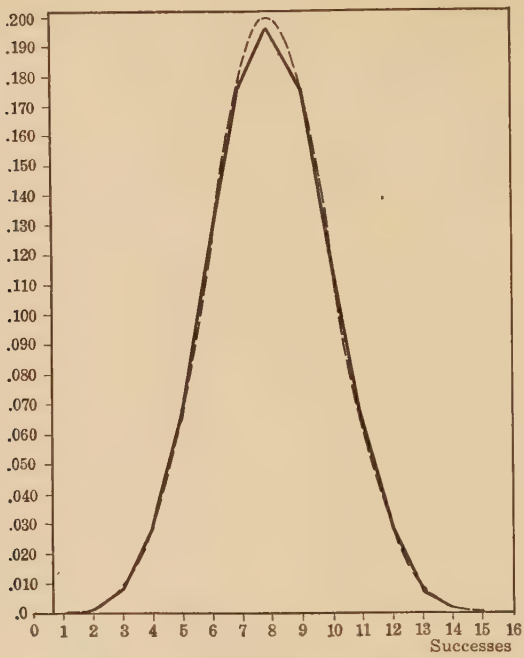


FIG. 54. The point binomial $(\frac{1}{2}+\frac{1}{2})^{16}$ compared with the normal curve

$$y=\frac{1}{\sqrt{8\pi}}e^{-\frac{x^2}{8}}$$

5. FITTING A NORMAL CURVE TO A FREQUENCY DISTRIBUTION OF DATA

The method for fitting a normal curve to a series of observations is similar to that just described, with the exception that areas and not ordinates are to be compared in determining the goodness of fit. The superposed curve is determined by taking its area, mean, and standard deviation equal to those obtained from the data.*

The work may be illustrated for the distribution of I.Q.'s given in Table 20 of Chapter VII. The necessary constants, already worked out, are

* For a more complete discussion of such fitting see Chapter XVI.

$N = 4834,$
$\sigma = 1.686 \times 10$ (1.661 with Sheppard's correction*),
$M = 89.28.$

Using formula (82), the equation of the desired normal curve will be

$$y = \frac{4834}{\sqrt{2\pi}(1.661)} e^{-\frac{1}{2}\left(\frac{x}{1.661}\right)^2}.$$

It will be noted that the standard deviation is expressed in units of class intervals, which is necessary in order to make y_0 comparable with the observed frequency in the interval at the mean, and bring the total area and frequency equal to N.

TABLE 45. NORMAL ORDINATES FOR I.Q. DATA

$\frac{x}{\sigma}$	SCALAR ABSCISSAS	z	$y = \frac{N}{\sigma} \times z$
0.0	89.28	.399	1161
± 0.5	97.58 and 80.98	.352	1024
± 1.0	105.89 and 72.67	.242	704
± 1.5	114.19 and 64.37	.130	378
± 2.0	122.50 and 56.06	.054	157
± 2.5	130.80 and 47.76	.018	52
± 3.0	139.11 and 39.45	.004	12
± 3.5	147.41 and 31.15	001	3
± 4.0	155.72 and 22.84	—	—

The value for y_0, when $x = 0$, is $\frac{4834}{2.5066 \times 1.661} = 1161$. From Table 42 it will be noted that the value for z at $x = 0$ is $\frac{1}{\sqrt{2\pi}} = .399$. It is therefore necessary to multiply this and all of the other ordinates taken from this table by the factor $\frac{N}{\sigma} = 2910$. Thus the ordinates at $\pm 0.5\,\sigma$ will have the values $2910 \times .352 = 1024$, $2910 \times .242 = 704$, etc.

The ordinates may be plotted at any convenient distances from the mean, say at multiples of 0.5σ, which must be worked

* See Chapter XVI, section 8.

out in actual scale units. The value for y_0 will, of course, be taken at 89.28, while the ordinates at $0.5\,\sigma$ will be located at $89.28 \pm .5(16.61)$, or at 97.58 and 80.98, etc. A complete list of values is shown in Table 45 on page 215.

A histogram of the observed frequencies and the fitted normal curve have been plotted on the same background in Fig. 55.

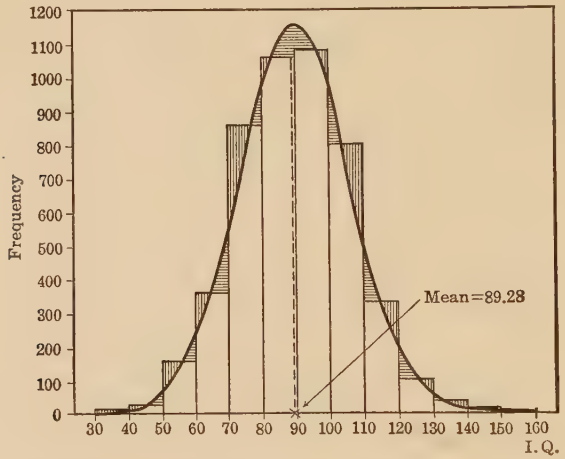


FIG. 55. Histogram for 4834 intelligence quotients with fitted normal curve

The agreement, as judged by mere inspection, appears to be rather good, but this method of comparison is worth very little in determining whether or not a particular mathematical curve adequately describes a body of data. The accurate method is to compare the discrepancies in frequency (area) between the histogram and the curve and determine whether or not the differences may or may not be accounted for by chance fluctuations of sampling. This test for goodness of fit will be applied in the chapter on Sampling (section 7).

6. Some Properties of the Normal Curve

From the fact that the normal curve is a continuous function it is now possible to find the probability for an occurrence between any two limits, x_1 and x_2. The actual frequency between these limits gives the number of favorable ways the event may happen, while the total frequency gives the total number of possible ways. The quotient of these two frequencies, or

$$\frac{\text{Frequency of occurrence between } x_1 \text{ and } x_2}{\text{Total frequency of all occurrences}},$$

then furnishes the desired measure of the probability.

In case the unit-area form of the normal curve is used, the denominator of this fraction becomes 1, and the probability for an occurrence between x_1 and x_2 is merely the area between these limits.

This area, which is known as the probability integral, may be found by using the appropriate values of $\frac{1}{2}\alpha$ given in Table 42 or in more extended tables such as Pearson's.

To illustrate the use of Table 42 in this connection, let us find the probability for an occurrence between 1σ and 2σ. This is represented in Fig. 56 by the shaded area. From the table the area from $x = 0$ to $x = 2$ is found to be .477, while the area from $x = 0$ to $x = 1$ is .341. The required area and probability is therefore the difference between these two values, or .136.

The same reasoning may be applied in the case of a distribution of observed data such as the 4834 I.Q.'s. In order to find the probability of getting an I.Q. between 130 and 140 in such a group it is only necessary to divide 36 (the number of favorable occurrences) by 4834 (the number of equally likely occurrences), and obtain .0074 as the required probability. Thus, if the 4834 I.Q.'s were recorded on little tickets and mixed up in a box, the chance of drawing a card with I.Q. between 130 and 140 would be .0074, or less than one in a hundred.

The probability integral is also useful in determining the chances that an occurrence will lie within or without a given middle range about the mean. Thus the probability (from Table 42) for an event between -3σ and $+3\sigma$ is the value of α at $x = 3$, that is, $2 \times .499$, or .998, while the probability for an occurrence beyond these limits in either direction is .002. By more extended tables,* these two values are .9973002 and

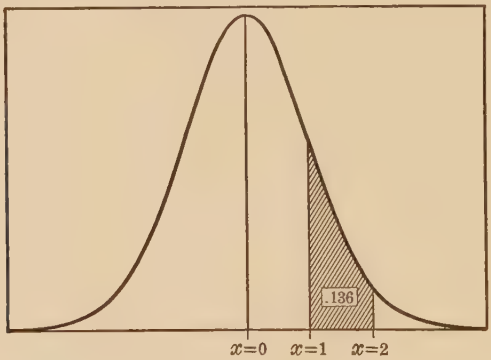


FIG. 56. Illustrating the area between $x = 1$ and $x = 2$ on a normal curve

.0026998, respectively. The probability for an occurrence beyond $\pm 6\sigma$ is .000000002, or only twice in a billion trials.

In case the probable error is used as a unit of measurement it is possible to determine the probabilities for an occurrence between the given multiples of *P.E.* when laid off on either side of the mean. Thus the chance of a deviate within ± 1 *P.E.* is $\frac{1}{2}$ (by definition). A short table of such probabilities is given on page 219.

Another interesting property of the normal curve makes it possible to find the mean of the portion between any two ordinates. Let the equation of the curve be taken in the form

$$z = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left\{ \begin{array}{l} \text{Ordinate of the normal curve, with} \\ \text{unit area and standard deviation} \end{array} \right\} \quad (83)$$

* Pearson, Tables for Statisticians and Biometricians. Cambridge University Press.

TABLE 46. PROBABILITIES THAT A DEVIATE WILL LIE WITHIN CERTAIN LIMITS ON A NORMAL CURVE

P.E.	PROBABILITY OF AN OCCURRENCE WITHIN A RANGE OF ± A GIVEN MULTIPLE OF *P.E.*
.5	.264
1.0	.500
1.5	.688
2.0	.822
2.5	.908
3.0	.957
3.5	.982
4.0	.993
4.5	.998

with unit area and standard deviation; let z_1 and z_2 be the ordinates at any two points x_1 and x_2, the second abscissa having the larger value; let ${}_1n_2$ be the area between these ordinates; and let ${}_1\bar{x}_2$ denote the mean of the inclosed portion. Then it may be proved * that

$$ {}_1\bar{x}_2 = \frac{z_1 - z_2}{{}_1n_2}. \quad \left\{ \begin{array}{l} \text{Mean of a portion of a normal curve,} \\ \text{with unit area and standard deviation} \end{array} \right\} \quad (85) $$

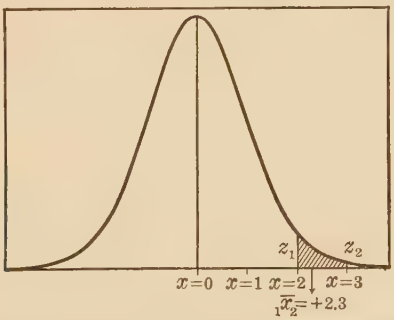


FIG. 57. Illustrating the mean of a portion of a normal curve between $x = 2$ and $x = 3$

* For any continuous function, $z = f(x)$, the mean between the limits x_1 and x_2 is given by $\int_{x_1}^{x_2} xzdx \Big/ \int_{x_1}^{x_2} zdx$. In the present case, $z = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and $\int_{x_1}^{x_2} zdx = {}_1n_2$. The integral in the numerator may be readily evaluated, giving $\Big[-z \Big]_{x_1}^{x_2}$, or $z_1 - z_2$. Therefore ${}_1\bar{x}_2 = \frac{z_1 - z_2}{{}_1n_2}$.

This theorem may be illustrated by finding the mean of the piece included between ordinates at $x = 2$ and $x = 3$. From Table 42, $z_1 = .054$ and $z_2 = .004$. The value for ${}_1n_2$ may be found by subtracting $\frac{1}{2}\alpha$ for x_1 from $\frac{1}{2}\alpha$ for x_2, that is to

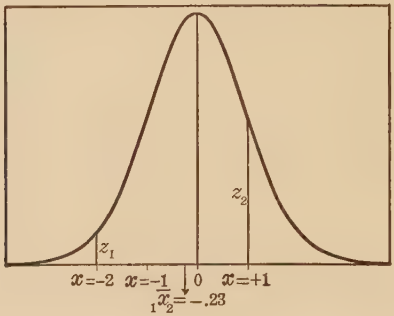


FIG. 58. Illustrating the mean of a portion of a normal curve between $x = -2$ and $x = +1$

say, $.499 - .477 = .022$ gives the area ${}_1n_2$ between the two ordinates. The required mean for this piece is therefore

$$ {}_1\bar{x}_2 = \frac{.054 - .004}{.022} = +2.3 \text{ (Fig. 57)}, $$

or 2.3 standard deviations above the mean of the whole curve. With Pearson's tables we find

$$ {}_1\bar{x}_2 = \frac{.0539910 - .0044318}{.0214002} = 2.31583. $$

It should be noted that ${}_1n_2$ is always positive, and that the sign of ${}_1\bar{x}_2$ is determined by the difference between the ordinates, which must be subtracted in the order indicated. Thus the mean of the piece between $x = -2$ and $x = 1$ will be obtained by adding the two values for $\frac{1}{2}\alpha$ and subtracting the larger from the smaller ordinate, that is, from Table 42,

$$ {}_1\bar{x}_2 = \frac{.054 - .242}{.818} = \frac{-.188}{.818} = -0.23 \text{ (Fig. 58)}. $$

7. Representing Data on a Normal Scale

In case we are dealing with a series of observations with standard deviation σ, and total frequency N, formula (85) may be modified so that the inclosed area is a fraction of the total, and the mean is expressed in units of the standard deviation, that is,

$$\frac{{}_1\bar{x}_2}{\sigma} = \frac{z_1 - z_2}{\frac{{}_1f_2}{N}} \cdot \left\{ \begin{array}{c} \text{Mean of a portion} \\ \text{of a normal curve,} \\ \text{with area} = N \end{array} \right\} \quad (86)$$

By means of the above formula it is now possible to represent a qualitative series of observations on a normal scale, assigning to each class the numerical value given by the mean of each sub-group. In this way the qualitative series has been converted into a quantitative one, the assumption being that the law behind the data is the normal distribution. This method is of the greatest importance because it makes possible the application of many formulas requiring numerical values for the classes (see Chapter XIV).

Any other curve might be used to represent such data, but as indicated at the beginning of this chapter the normal curve is the best single approximation to most educational data, and very fortunately it is extremely simple to apply.

As an example, let us represent the following qualitative series on a linear and then on a normal scale. The data are general health estimates of school children made by several physicians.

Table 47. Health Data with Percentage Frequencies

Health of Child	f	Percentage f
Very robust	16	2.0
Robust	199	24.4
Normal	345	42.3
Rather delicate	115	14.1
Delicate	124	15.2
Very delicate	16	2.0
	815	100.0

If we assume that the attribute, health, is distributed with equal frequency along a scale, the resulting series will form a long rectangle. The mean of each piece, occurring at the middle, might then be taken as a numerical measure of the class. This

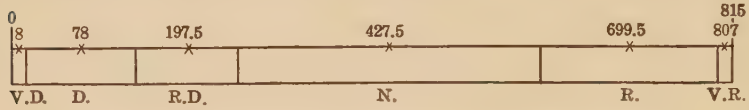


FIG. 59. Rectangular distribution of the health series

method, however, would be unsound because it assumes a form of distribution totally unlike any observed for such traits.

Assuming that health is normally distributed, the series may be represented as in Fig. 60. It is now possible to determine the means of the various pieces by the use of Table 43. The need of such a table becomes apparent when it is noted that areas and not deviates are furnished by the data. While it is better to use more extended tables, such as Kelley's* or Holzinger's, the work will be illustrated by Tables 48 and 49, the figures in parentheses being obtained from Holzinger's Table XII.

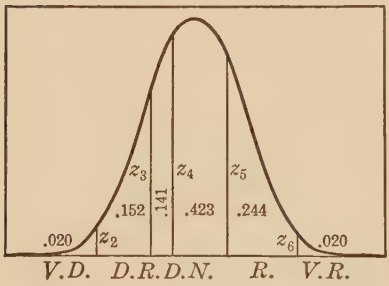


FIG. 60. Representation of the health data on a normal scale

If the ordinates are designated as $z_1, z_2, z_3 \cdots z_7$ it is evident that z_1 and z_7 are zero. The other five ordinates, inclosing various pieces, may be obtained by reducing the areas to total unit area, and entering Table 43 with the proper value of $\frac{1}{2}\alpha$. Thus the area to the right of z_6 is .020, so that $\frac{1}{2}\alpha = .480$; the area to the right of z_5 is .264, giving $\frac{1}{2}\alpha = .236$; while the area to the right of z_4 is .687, for which $\frac{1}{2}\alpha = .187$, as shown in Table 48.

* T. L. Kelley, Statistical Method. The Macmillan Company.

TABLE 48. SHOWING THE CALCULATION OF THE FIVE ORDINATES FOR THE HEALTH DATA REPRESENTED ON A NORMAL SCALE

ORDINATE	AREA BETWEEN ORDINATES	$\frac{1}{2}\alpha$	VALUE OF ORDINATE
z_7		.50 (.500)	.000
	.020		
z_6		.48 (.480)	.048 (.0484)
	.244		
z_5		.24 (.236)	.324 (.3269)
	.423		
z_4		.19 (.187)	.353 (.3543)
	.141		
z_3		.33 (.328)	.253 (.2550)
	.152		
z_2		.48 (.480)	.048 (.0484)
	.020		
z_1		.50 (.500)	.000

The means may now be obtained by subtracting the proper ordinates and dividing by the area between them. The work may then be set down as follows:

TABLE 49. SHOWING THE CALCULATION OF THE MEANS OF THE HEALTH CATEGORIES

MEAN	VALUE FROM 3-PLACE TABLE	VALUE FROM 4-PLACE TABLE
$\frac{6\bar{x}_7}{\sigma}$	$\frac{.048 - .000}{.020} = +2.40$	$\frac{.0484 - .0000}{.020} = +2.42$
$\frac{5\bar{x}_6}{\sigma}$	$\frac{.324 - .048}{.244} = +1.13$	$\frac{.3269 - .0484}{.244} = +1.14$
$\frac{4\bar{x}_5}{\sigma}$	$\frac{.353 - .324}{.423} = +0.07$	$\frac{.3543 - .3269}{.423} = +0.06$
$\frac{3\bar{x}_4}{\sigma}$	$\frac{.253 - .353}{.141} = -0.71$	$\frac{.2550 - .3543}{.141} = -0.70$
$\frac{2\bar{x}_3}{\sigma}$	$\frac{.048 - .253}{.152} = -1.35$	$\frac{.0484 - .2550}{.152} = -1.36$
$\frac{1x_2}{\sigma}$	$\frac{.000 - .048}{.020} = -2.40$	$\frac{.0000 - .0484}{.020} = -2.42$

As a check on the computation, the products of the means by the corresponding areas, when added, should equal zero (the mean of the whole distribution), for example,

$$2.40 \times .020 + 1.13 \times .244 + 0.07 \times .423$$
$$- 0.71 \times .141 - 1.35 \times .152 - 2.40 \times .020 = + .00002.$$

The check in this case is accidentally close.

According to these results the "health" difference between a typical robust and a typical normal child is $1.06\,\sigma$, while the difference between a normal and a rather delicate child is $0.78\,\sigma$. It will also be noted that the mean of the normal health group is very close to zero (the mean of the whole distribution) and that the very delicate and very robust groups are equally divergent from this point.

While comparisons such as these are often of great value in analyzing a body of qualitative data, the chief use of this scaling method is in studying the relationship between several traits. It is possible, for example, to obtain a measure of the relationship (correlation) between health and general nutrition, or between health and intelligence, by representing the pairs of characters on normal scales (see Chapter XIV).

8. The Scaling of Test Questions

The normal curve has been widely used in the scaling of standardized test questions. Assuming that the ability of the pupils is measured by the difficulty of the exercises, the latter may be represented on a normal scale. If nearly all of a group of pupils solve a problem, its value will be low; if 50 per cent do an exercise correctly, its value will be at the mean; while if very few succeed on an item, it will be located high on the normal scale. The particular scale value of the item is thus determined by the per cent of the group solving the problem correctly.

In Fig. 61 the percentage of correct solutions is shown by the shaded area, and the value of the item is given by the corresponding abscissa. In order to obtain this value for this example it is only necessary to enter Table 43 with $\frac{1}{2}\alpha = .20$, giving $x = .524$. The problem thus has a difficulty or ability value of .524 standard deviation above the mean.

By taking the mean at zero it will be noted that negative values of the deviates will occur. This may be overcome by shifting the origin to some convenient point, say 5σ below the mean, as shown in Fig. 61. Such an arbitrary origin should not be confused with the point for "just no ability in the trait" sought after by some test makers. Just as temperature is measured on the Fahrenheit scale from an arbitrary zero, not representing the

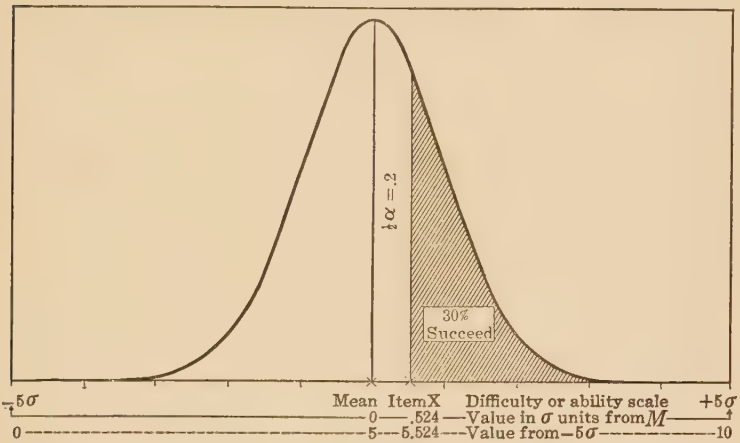


FIG. 61. Illustrating the scaling of test questions with the normal curve

point for no heat, so educational scales may be taken from any convenient reference point, not representing "just no ability."

It is possible to scale the items one at a time, or several at once, as proposed by McCall.* The procedure by the first method may be further illustrated with some reading questions given to a large group of twelve-year-old pupils. In Table 50 the first and sixth questions will have negative deviates given by entering Table 43 with $\frac{1}{2}\alpha = (.98 - .50)$ and $(.75 - .50)$, while the other two questions will have positive deviates, being at the right of the mean. The final scaled values are obtained by merely adding 5 to each of these deviates.

* McCall, How to Measure in Education. The Macmillan Company.

TABLE 50. SHOWING A METHOD FOR SCALING EACH TEST ITEM

PROBLEM	PER CENT OF PUPILS ANSWERING CORRECTLY	$\frac{1}{2}\alpha$	$\frac{x}{\sigma}$	SCALED VALUE $= \frac{x}{\sigma} + 5$
1	98	.48	− 2.054	2.946
6	75	.25	− 0.674	4.326
13	46	.04	+ 0.100	5.100
24	4	.46	+ 1.751	6.751

By McCall's method it is necessary to note the percentage of successful replies to at least 0, 1, 2, 3, · · · questions, the items being previously arranged in rough order of difficulty. Thus, with the above reading material, the following results were obtained:

TABLE 51. SHOWING MCCALL'S METHOD OF SCALING TEST QUESTIONS

NUMBER OF QUESTIONS CORRECT = Q	NUMBER OF PUPILS OBTAINING GIVEN Q	PERCENTAGE OF PUPILS EXCEEDING, PLUS HALF THOSE REACHING Q	$\frac{1}{2}\alpha$	$\frac{x}{\sigma}$	SCALED VALUE $= \frac{x}{\sigma} + 5$
0	1	99.9	.499	−3.090	1.910
1	3	99.5	.495	−2.576	2.424
2	5	98.6	.486	−2.197	2.803
3	7	97.3	.473	−1.927	3.073
4	9	95.6	.456	−1.706	3.294
–	–	—	—	—	—
–	–	—	—	—	—
–	–	—	—	—	—
Total	462				

In order to obtain the percentage of pupils above a given class value, McCall has added one half of the number of pupils at Q to the number exceeding Q, and then divided by the total number in the sample. The arithmetic for the first two values in the above table will then be

$$\frac{461 + \frac{1}{2} \times 1}{462} = .999, \qquad \frac{458 + \frac{1}{2} \times 3}{462} = .995.$$

The deviates and scaled values are obtained by Holzinger's Table XII and by adding 5 to eliminate the negative signs. McCall, however, multiplies these last values by 10 and calls them T scores.

According to the first method of scaling, the score of a pupil answering the first four questions correctly would be the sum of the four scaled values. By McCall's method, such a performance would be scaled by assigning the T score corresponding to $Q = 4$ from Table 51. McCall's method is, therefore, very convenient, but there is some doubt as to the assumption that different sequences of problems (for example, 1, 2, 3, 4, 5, · · ·, 1, 2, 4, 5, 6, etc.) obtained by various pupils have the same value.

It should be noted that great precision in scaling test material is idle. The figures above have been put down as they came from the tables, but they should ordinarily be rounded off to one decimal place at most.

Scaled values are often an unnecessary refinement in measuring large groups as evidenced by the high correlations between scaled and unscaled items. Professor Douglass,* for example, found a correlation of about .98 between weighted and unweighted algebra scores, a result which is much higher than the reliability of the tests themselves. He concluded that the unscaled values give the relative standing of the pupil with sufficient accuracy for ordinary testing uses.

In the case of individual measurements, scaled values also lose much of their significance because they are based upon a large group and may not apply to a single person. Thus for the whole group, problem 1 in Table 50 has the value 2.9, while problem 6 has the value 4.3. For a given individual, however, it is not improbable that the two items are equally difficult.

* H. R. Douglass and P. L. Spencer, "Is it Necessary to weight Exercises in Standard Tests?", *Journal of Educational Psychology*, February, 1923, p. 109. Dr. Scates and the writer have also found correlations of .994, .995, .997, and .998 between weighted and unweighted scores, the number of items weighted varying from six to ten, and the weights being quite different.

The chief advantage in scaling by the above methods is that test results are thereby expressed in comparable units from comparable reference points (for example, a T of 60 on any test means 1 σ above the mean). Weighted values may also be used to graduate test material in order of difficulty or to arrange parallel groups of items such as spelling words of equal difficulty.

When test material is to be scaled by the judgment of experts rather than by the performance of the pupils, the normal curve may again be employed. The procedure here is to have the judges arrange the pupil specimens (say drawings) in order of merit according to their best opinion. If 50 per cent of the judges rate specimen A as better than specimen B, these two are regarded as of equal value on the assumption that "equally rated differences are equal unless they are always or never noticed."*

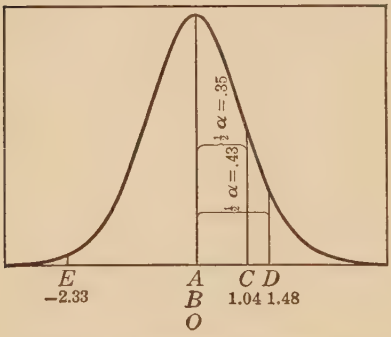


FIG. 62. Illustrating the scaling of items for a product test

If 85 per cent of the judges rate specimen C as better than A, then the difference in value between A and C is obtained by finding the deviate for $\frac{1}{2}\alpha = (.85-.50) = .35$. Thus in Fig. 62, C has the scaled value $x = 1.04$, the unit being the standard deviation with the origin at the mean. If the percentage of judges rating C better than B is 83, then a new scaled value $x = .95$ may be averaged with 1.04 etc.

By calculating similar differences for all pairs of specimens a series of scaled values is obtained. The origin may be taken at an arbitrary point (such as -5σ), but is often selected at the specimen which most judges consider worthless.

* This is known as the Cattell-Fullerton Theorem.

As a further illustration of the arithmetic, two items may be added to the above series. Problem *D* is rated better than *A* by 93 per cent of the judges, while problem *E* is regarded by most as worthless. Assuming that 99 per cent of the judges rate *A* better than *E*, the value of the latter becomes -2.33. The scaled values may now be written as follows:

Origin	Values of Specimens			
	E	*A*, *B*	*C*	*D*
A	−2.33	0	1.04	1.48
-5σ	2.67	5	6.04	6.48
E	0	2.33	3.37	3.81

The last row of numbers is probably most convenient to use, but zero is then only a rough approximation to "just no ability."

All these results were obtained by using *A* as the item of comparison, but approximately the same values would have been secured if all differences had been computed with reference to problem *E*. In the final scale it is usually best to select only those items which differ from one another by fairly large amounts (say $.5\,\sigma$), because, in using the scale, finer differences cannot be readily noted.

EXERCISES

1. Find the probabilities of occurrences within the following ranges for a normal curve. Use Holzinger's Table XI.

Range	Probability (*Ans.*)
$-2.5\,\sigma$ to $-1.5\,\sigma$	.0606
$-2.5\,\sigma$ to $+2.5\,\sigma$	.9876
$+1.0\,\sigma$ to $+3.0\,\sigma$	.1574
$+3.54\,\sigma$ to $+3.88\,\sigma$	.0001
$-0.62\,\sigma$ to $+2.79\,\sigma$	.7298

2. Allowing a range of $1.2\,\sigma$ for each of the five marks *A*, *B*, *C*, *D*, and *E*, find the percentages of such marks under a normal distribution. (3.46, 23.84, 45.14, 23.84, 3.46. *Ans.*)

3. Represent the following data on a normal scale and find the means of the five categories:

Grade of School Work	Percentage Frequency	Mean (*Ans.*)
A	5	$+2.062\,\sigma$
B	21	$+1.054\,\sigma$
C	49	$-0.013\,\sigma$
D	18	$-1.019\,\sigma$
E	7	$-1.919\,\sigma$

4. Eighty-eight per cent of a group of judges rate drawing *A* as better than drawing *B*, while 75 per cent rate *B* as better than *C*. Assuming that 99 per cent of the judges have rated *C* as better than *X*, which has no merit whatsoever, obtain the values of the drawings *C*, *B*, and *A* with respect to *X*.

($X = 0$; $C = 2.3263\,\sigma$; $B = 3.0008\,\sigma$; $A = 4.1758\,\sigma$. *Ans.*)

5. In a large group of children, the percentage of those who solved a given example, with five specified examples considered one at a time, varied as follows: 94, 87, 61, 43, 11. Find the σ value of each example, using as origin a point $5\,\sigma$ below the mean.

(3.4452, 3.8736, 4.7207, 5.1764, 6.2265. *Ans.*)

6. Find the percentage distribution of five marks, using a range of $1\,\sigma$ for each. (6.06, 24.17, 38.30, 24.17, 6.06. *Ans.*)

7. Verify the following results:

Number of Questions Correct = *Q*	Percentage of Pupils obtaining Given *Q*	Per Cent exceeding, plus Half those reaching *Q*	*T* Score (*Ans.*)
0	2	99	26.7
1	6	95	33.6
2	12	86	39.2
3	18	71	44.5
4	20	52	49.5
5	14	35	53.9
6	12	22	57.7
7	10	11	62.3
8	6	3	68.8

8. Calculate the ordinates for $(.5 + .5)^9$ and compare them with those of the normal curve.

9. Fit normal curves to the distributions of I.Q.'s given in Table 55 of Chapter XIII. Use columns 1, 2, 3, and 4.

CHAPTER XIII

SAMPLING AND RESPONSE ERRORS

1. Introductory

All statistical quantities such as averages and measures of relationship are based upon samples. The results found from one sample will never quite agree with those found from another, nor with those from the whole population from which the samples were chosen. In determining the stability of a given measure or in comparing the results from different groups it is therefore important to know the probable extent of such fluctuations.

Thus a correlation of .30 may appear to indicate some relationship between two traits, but if on taking another sample the coefficient is found to be .10, we can place little confidence in either of the two results. Some measure of the likely variation from sample to sample is clearly desirable.

Again, in the case of a control experiment, two means might be obtained for comparison, their difference being the test of the relative superiority of two methods of learning. For example, the mean gain might be 22 for a control group, and 20 for a practice group. The difference is 2, but whether or not it is of any significance remains to be shown. It might be that by repeating the experiment the difference would come out to be -3 in favor of the other group. Here also a critical test of such differences under sampling is necessary.

The stability of a statistical constant from sample to sample is often called its *reliability** and is measured by the use of sampling formulas to be discussed in the present chapter. On

* This term should not be confused with the *reliability coefficient* r_{11} for a test. It might therefore be better to use the expression "sampling reliability" for the former.

account of the rather elaborate mathematics involved only a few of the proofs of these formulas will be given, but their use and interpretation as applied to a variety of educational problems will be treated at some length.

Sampling formulas as applied to statistical data are usually approximations, their accuracy depending on certain assumptions in the proofs and especially upon the number of cases involved. The chief danger in using such formulas without being familiar with the proofs may be avoided by never applying them to a small number of cases (say less than thirty).

In the last section of this chapter some of the current formulas for dealing with response errors will be presented. As noted in Chapter V, response errors are due to the variability of performance within the individual measured or tested.

2. Sampling Error in the Mean

If the true mean of an indefinitely large number of observations be denoted by M and their standard deviation by σ_x, and if the mean of a randomly drawn sample of N individuals be represented by M_1, the difference $M - M_1$ is known as the *sampling error* in the mean. It can be shown theoretically that if repeated samples of N be randomly drawn from the population, the differences $M - M_1$ will be distributed around zero with a standard deviation given by the formula

$$\sigma_M{}^* = \frac{\sigma_x}{\sqrt{N}}. \quad \left\{\begin{matrix}\text{Standard error} \\ \text{of the mean}\end{matrix}\right\} \qquad (87)$$

If the size of the samples is large the distribution of $M - M_1$ tends to follow a normal curve even though the population sampled is not normal.

* A good proof of this formula is given in Jones's "First Course in Statistics," p. 153. G. Bell & Sons, Ltd., London, 1921. The reasonableness of the formula is at once apparent from the fact that a small dispersion and a large number of cases decrease the size of σ_M.

As an approximation to an indefinitely large number of cases let us assume that we have 50,000 observations of a certain variable with the mean equal to M, and that samples of 500 be drawn. The means of these samples, which may be denoted by M_1, M_2, M_3, . . . M_{100}, will be distributed about M in a frequency curve resembling that which would have been found had the number of samples been increased indefinitely. A hypothetical distribution of such means is shown in Fig. 63. The mean of all the samples is 148, and the standard deviation, σ_M, is 1.71.

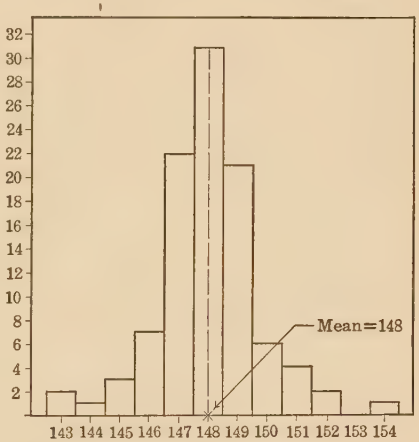


FIG. 63. Hypothetical distribution of means from one hundred samples

Now let us assume that one of the samples of 500 cases furnishes a mean M_1 equal to 146, and a standard deviation σ_{x_1} equal to 37.12. By substituting these values in formula (87) we then find $\sigma_M \doteq 1.66$. If the means and standard deviations from other samples had been used in this formula, very nearly the same results would have been obtained for σ_M because σ_x will vary but slightly from sample to sample, provided the size of the sample is large.

It thus appears that in dealing with only one sample the mean of the whole population is unknown, but may be approximated by M_1, and that the formula $\sigma_{M_1} = \frac{\sigma_{x_1}}{\sqrt{N}}$ gives the best obtainable approximation to the true standard deviation σ_M.

The probable error of the mean is given by the formula

$$P.E._M = .6745 \frac{\sigma_x}{\sqrt{N}}. \quad \left\{ \begin{matrix} \text{Probable error} \\ \text{of the mean} \end{matrix} \right\} \quad (88)$$

If the true values for M and σ_x were known it would then be possible to find a range on the normal scale within which it is almost certain that an observed mean M_1 must lie. In actual practice, however, it is M_1 and not M that is known, so that this argument must be reversed.

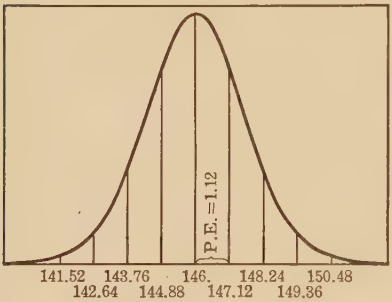


FIG. 64. Illustrating various ranges of probable error on a normal curve

The theoretical curve in Fig. 64 represents the inverse probability for various positions of the true mean when M_1 is known. The value for $P.E._{M_1}$ by formula (88) is $.6745 \times 1.66 = 1.12$. Since half the area of the curve lies between $M_1 - P.E.$ and $M_1 + P.E.$, or between 144.88 and 147.12, the probability that the true mean lies between these limits is .5, and the result is ordinarily written $M_1 = 146 \pm 1.12$. By similar argument we find that the chances are over 99 in 100 that the true mean will lie in the range $M_1 \pm 4\ P.E.$, or between 141.52 and 150.48 as shown in Table 53. This range is the usually accepted zone of safety.

TABLE 52. PROBABILITIES THAT THE TRUE MEAN WILL LIE WITHIN A GIVEN RANGE

RANGE	PROBABILITY THAT M LIES WITHIN GIVEN RANGE
$M_1 \pm 1\ P.E.$ (144.88–147.12)	.500
$M_1 \pm 2\ P.E.$ (143.76–148.24)	.822
$M_1 \pm 3\ P.E.$ (142.64–149.36)	.957
$M_1 \pm 4\ P.E.$ (141.52–150.48)	.993
$M_1 \pm 5\ P.E.$ (140.40–151.60)	.999

The calculation of probable errors of the mean given by formula (88) is facilitated by the use of tables giving the values

of $\chi_1 = \frac{.6744898}{\sqrt{N}}$. The probable error is obtained by multiplying the observed value of σ_x by the tabled value of χ_1. Thus, for $\sigma_x = 13.1$ and $N = 147$, we find from Holzinger's Tables for Students, Table IX, that $\chi_1 = .0556$. The value for $P.E._M$ is therefore $.0556 \times 13.1 = .728$.

3. The Probable Error of the Difference between Two Means

One of the most useful formulas in sampling is that for testing whether or not small differences may have arisen from chance. The formula may be employed with a variety of statistical measures, but is most frequently applied in the case of the mean.

If the variables in two groups, and hence their means, are quite independent of one another the probable error of the difference $M_1 - M_2$ is given by the formula

$$P.E._{M_1-M_2} = \sqrt{(P.E._{M_1})^2 + (P.E._{M_2})^2}. \left\{\begin{array}{c}\text{Probable error of the}\\ \text{difference between two}\\ \text{uncorrelated means}\end{array}\right\} \quad (89)$$

The use of this formula may be illustrated in the case of a control experiment in the teaching of physics. Two groups of pupils were equated with respect to intelligence and initial ability in a type of high-school physics. After teaching one group by the lecture method and the other group by the demonstration method a final test was given and results found as shown in Table 53.

Table 53. Data from Physics-Teaching Experiment

	Lecture Group	Demonstration Group
Population	$N_1 =$ 37	$N_2 =$ 41
Mean intelligence score	137	138
Mean score on initial physics test	74.3	74.3
Mean score on final physics test	$M_1 =$ 91.43	$M_2 =$ 89.64
Standard deviation for final physics test	$\sigma_1 =$ 7.08	$\sigma_2 =$ 7.23
Probable error of M	$P.E._{M_1} =$ .785	$P.E._{M_2} =$ .761

The probable errors of the means are given by formula (88). Substituting these values in formula (89), we find that

$$P.\,E._{M_1-M_2} = \sqrt{(.785)^2 + (.761)^2} = 1.09,$$

the arithmetic being quickly done with a table of squares.

The difference between final scores may now be written

$$M_1 - M_2 = 91.43 - 89.64 = 1.79 \pm 1.09.$$

Such a difference is regarded as *insignificant*, or such that it is not unlikely that the true difference is zero. This is illustrated in Fig. 65. Speaking approximately, since the number of observations is small, the probability that the true difference lies outside a range of $\pm 2\,P.E.$, or $-.39$ to .397, is about .18 by Table 54. The probability that the true difference will be outside the range 0 to 3.58 may be had from Table 54 by entering with $\frac{x}{P.E.} = \frac{1.79}{1.09} = 1.64$, the result being approximately .27. The chances are, therefore, approximately one in four that the true difference will be as small as 0 or as large as 3.58.

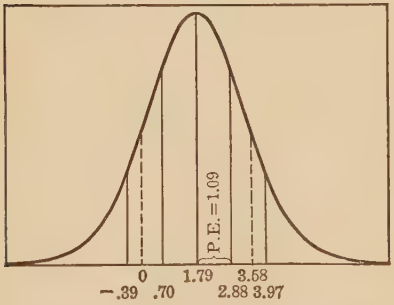


FIG. 65. Illustrating the probability that an observed difference will be as low as zero or as high as 3.58

In view of the above test the whole study is to be regarded as inconclusive. We have no right to ascribe the observed difference of 1.79 to the superiority of the lecture method when it can be readily accounted for by chance fluctuations in sampling. It should also be noted that there are a large number of variable factors to be controlled in such an experiment. These factors can never be perfectly controlled and will undoubtedly affect the final result to some extent. It is assumed that the errors in sampling are independent of these factors.

TABLE 54. PROBABILITIES OF THE OCCURRENCE OF DEVIATIONS RELATIVE TO THE SIZE OF THE PROBABLE ERROR

$\frac{x}{P.E.}$	PROBABILITY OF A DEVIATION BEYOND $\pm \frac{x}{P.E.}$	$\frac{x}{P.E.}$	PROBABILITY OF A DEVIATION BEYOND $\pm \frac{x}{P.E.}$
1.0	.5000	3.0	.0430
1.1	.4581	3.1	.0365
1.2	.4183	3.2	.0309
1.3	.3806	3.3	.0260
1.4	.3450	3.4	.0218
1.5	.3117	3.5	.0182
1.6	.2805	3.6	.0152
1.7	.2515	3.7	.0126
1.8	.2247	3.8	.0104
1.9	.2000	3.9	.0085
2.0	.1773	4.0	.0070
2.1	.1567	4.1	.0057
2.2	.1378	4.2	.0046
2.3	.1208	4.3	.0037
2.4	.1055	4.4	.0030
2.5	.0918	4.5	.0024
2.6	.0795	4.6	.0019
2.7	.0686	4.7	.0015
2.8	.0589	4.8	.0012
2.9	.0505	4.9	.0009

The general rule, already noted, is that a difference or a statistical constant of any sort is not significant unless it is at least four times its probable error.

Table 54 gives the probabilities for deviations greater than $\frac{x}{P.E.}$ and less than $-\frac{x}{P.E.}$ for various values of $\frac{x}{P.E.}$, that is, the fraction of the area under a normal curve beyond these limits.

4. THE PROBABLE ERRORS OF CERTAIN CONSTANTS FOR A NORMAL DISTRIBUTION

The probable error of the mean may be used for any form of distribution, but in the case of certain other constants, it is assumed in the proofs that the distribution is normal. The following formulas should therefore be used only in case the

observed distribution from which the constants are obtained approximates the normal probability curve.

The probable error of the median is given by the formula

$$P.E._{Md} = \frac{.84535\,\sigma_x}{\sqrt{N}} = 1.2533\,P.E._M. \quad \left\{\begin{matrix}\text{Probable error}\\ \text{of the median}\end{matrix}\right\} \quad (90)$$

Inasmuch as the sampling error in the median is about 25 per cent more than in the mean, the greater reliability of the latter is at once apparent. For certain very peaked (leptokurtic) distributions the median may be more reliable,* but for the large majority of problems the distributions are roughly normal and the mean is to be preferred.

The standard deviation is one of the most reliable of all statistical constants, its probable error being given by

$$P.E._\sigma = \frac{.6745\,\sigma}{\sqrt{2N}} = \frac{.4769\,\sigma}{\sqrt{N}} = .7071\,P.E._M. \quad \left\{\begin{matrix}\text{Probable error of the}\\ \text{standard deviation}\end{matrix}\right\} \quad (91)$$

In case $P.E._M$ is also required, the last form on the right is probably the most convenient for computation.

The coefficient of variation, V, has for its probable error

$$P.E._V = \frac{.6745\,V}{\sqrt{2N}}\left\{1 + 2\left(\frac{V}{100}\right)^2\right\}^{\frac{1}{2}}. \quad \left\{\begin{matrix}\text{Probable error of the}\\ \text{coefficient of variation}\end{matrix}\right\} \quad (92)$$

The calculation is facilitated by the use of Pearson's Tables V and VI, which give the values of χ_2 and ψ. The formula may then be written

$$P.E._V = \left\{\frac{.6745}{\sqrt{2N}}\right\}\left\{V\left[1 + 2\left(\frac{V}{100}\right)^2\right]^{\frac{1}{2}}\right\} = \chi_2\psi. \quad \left\{\begin{matrix}\text{Probable error}\\ \text{of } V \text{ with Pear-}\\ \text{son's Tables}\end{matrix}\right\} \quad (93)$$

The probable error of the correlation coefficient is

$$P.E._r = \frac{.6745\,(1 - r^2)}{\sqrt{N}}. \quad \left\{\begin{matrix}\text{Probable error† of the}\\ \text{correlation coefficient}\end{matrix}\right\} \quad (94)$$

* Yule, Introduction to Statistics, p. 338.

† The student is warned that this formula should not be applied when N is small and at the same time r is large. Misleading results may follow for such cases, as $N = 20$ and $r = .5$, $N = 50$ and $r = .8$, or $N = 100$ and $r = .9$.

Complete tables* for this error have been worked out by the writer for every value of N from 20 to 100, and by tens thereafter up to 1000. A shorter table is also found in Table X of Holzinger's Statistical Tables for Students.

An approximate value for the probable error of the correlation ratio is

$$P.E._{\eta} = \frac{.6745\,(1-\eta^2)}{\sqrt{N}}, \quad \left\{\begin{array}{c}\text{Probable error of the}\\ \text{correlation ratio}\end{array}\right\} \tag{95}$$

so that the above tables may also be used for this measure of association.

In the case of the regression coefficients $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ and $b_{yx} = r\frac{\sigma_y}{\sigma_x}$, the probable errors are

$$P.E._{b_{xy}} = .6745\frac{\sigma_x}{\sigma_y}\frac{\sqrt{1-r^2}}{\sqrt{N}} \tag{96a}$$

$$\text{and} \qquad P.E._{b_{yx}} = .6745\frac{\sigma_y}{\sigma_x}\frac{\sqrt{1-r^2}}{\sqrt{N}}. \quad \left\{\begin{array}{c}\text{Probable errors}\\ \text{of regression co-}\\ \text{efficients}\end{array}\right\} \tag{96b}$$

Similar formulas are applied in the case of partial regression coefficients (see Chapter XV); that is,

$$P.E._{b_{12\cdot k}} = .6745\,\frac{\sigma_{1\cdot 2k}}{\sigma_{2\cdot k}\sqrt{N}}, \quad \left\{\begin{array}{c}\text{Probable error of higher-order}\\ \text{regression coefficient}\end{array}\right\} \tag{97}$$

k being any collection of secondary subscripts other than 1 or 2. These last formulas should not be confused with formulas (45) and (46), which give the probable errors of estimate of a single score by the lines of regression.

In testing for linearity of regression, the probable error of $\delta = \eta^2 - r^2$ has already been used. The formula is

$$P.E._{\delta} = \frac{2(.6745)}{\sqrt{N}}\sqrt{(\eta^2-r^2)\{(1-\eta^2)^2-(1-r^2)^2+1\}}. \quad \{\text{Probable error of } \eta^2 - r^2\} \tag{98}$$

If $\eta^2 - r^2$ is to be less than three times its probable error, the above expression reduces to formula (67) of Chapter X.

* Karl J. Holzinger, Tables of the Probable Error of the Correlation Coefficient, Tracts for Computers No. XII, p. 35. Cambridge University Press, England, 1925.

5. Some Applications of Probable Error Formulas

One important use of the sampling theory is to determine whether or not two or more samples belong to the same or to different types of populations. This may be illustrated in the case of the distribution of 4834 intelligence quotients given in Table 20. The total distribution may be broken up into the sub-groups given in Table 55.

From the means and standard deviations at the bottom of the table, we may now test the difference between various groups designated from 1 to 6. If A and B are any two independent measures, formula 89 becomes

$$P.E._{A-B} = \sqrt{(P.E._A)^2 + (P.E._B)^2}.$$

Using this formula together with (88) and (91) we find:

$$M_1 - M_2 = -5.52 \pm \sqrt{(.27)^2 + (.34)^2} = -5.52 \pm .43$$

and $$\sigma_1 - \sigma_2 = 5.98 \pm \sqrt{(.19)^2 + (.24)^2} = 5.98 \pm .31,$$

both differences being clearly significant. The grade and high school city children are thus to be regarded as distinctly different intellectual types, the differences being probably due to selection.

By similar calculations we obtain:

$$M_1 - M_3 = 9.34 \pm .37, \qquad \sigma_1 - \sigma_3 = 1.07 \pm .26,$$
$$M_2 - M_3 = 14.86 \pm .42, \qquad \sigma_2 - \sigma_3 = 4.91 \pm .31.$$

Since all these differences are significant, the three white groups are to be considered as samples from essentially different types of populations.

The means for the two negro groups are found to be significantly lower than those for any of the white groups. The difference $M_4 - M_5$, $(2.64 \pm .78)$, does not prove to be significant by the usual test. From Table 54, however, it will be found that the odds are about 45 to 1 that city and country negroes are to be regarded as distinct intellectual types.

TABLE 55. DISTRIBUTIONS OF INTELLIGENCE QUOTIENTS FOR VARIOUS POPULATION TYPES

I.Q.	City Whites Grade Schools 1	City Whites High Schools 2	Country Whites 3	City Negroes 4	Country Negroes 5	All Groups 6
150–160	2	–	–	–	–	2
140–150	9	–	3	–	–	12
130–140	27	1	8	–	–	36
120–130	73	12	17	1	–	103
110–120	176	68	63	10	1	318
100–110	351	153	227	59	9	799
90–100	413	111	347	177	26	1074
80–90	280	42	447	253	37	1059
70–80	174	2	329	309	54	868
60–70	41	–	116	170	39	366
50–60	10	–	51	86	16	163
40–50	3	–	3	12	7	25
30–40	1	–	–	7	1	9
Total	1560	389	1611	1084	190	4834
Mean	96.76 ± .27	102.28 ± .34	87.42 ± .25	78.90 ± .29	76.26 ± .72	89.28 ± .16
S. D.	16.03 ± .19	10.05 ± .24	14.96 ± .18	14.39 ± .21	14.81 ± .51	16.86 ± .12

It is evident from the above comparisons that all five groups making up the total are to be regarded as samples from quite distinct population types. This lack of homogeneity doubtless accounts in part for the fact that group 6 does not furnish a good example of a normal curve.

Another application of the formula

$$P.E._{A-B} = \sqrt{(P.E._A)^2 + (P.E._B)^2}$$

may be made in the comparison of correlation coefficients. In the same number of the *Journal of Educational Psychology* two writers* presented correlations between mental ages on the Binet and the Herring intelligence tests. Dr. Herring gives the value $r = .987 \pm .002$, obtained from 116 twelve-year-old children, and Dr. Avery finds as his highest correlation, $r = .824 \pm .031$, from a group of 48 first-grade children. These two correlations are independent, since they were obtained from different groups. The difference by the above formula is then $.163 \pm .031$, which is more than five times its probable error, and therefore significant. A probable explanation† of the difference between these correlations lies in the fact that one of the tests is much more reliable than the other when applied to very young children.

In case the measures A and B are correlated the formula for testing the significance of the difference $A - B$ becomes

$$P.E._{A-B} = \sqrt{(P.E._A)^2 + (P.E._B)^2 - 2\,R_{AB}(P.E._A)(P.E._B)}, \quad (99)$$

{Probable error of difference with correlated measures}

where R_{AB} is the correlation between the sampling errors in A and B.

For two means M_1 and M_2 from correlated material, the correlation between the sampled means, $R_{M_1M_2}$, is equal to r_{12}, which is the correlation between the observed variables, so that

* John P. Herring, "Reliability of the Stanford and the Herring Revision of the Binet-Simon Tests," and A. T. Avery, "Comparison of Stanford and Herring Revisions Given to First-Grade Children," *Journal of Educational Psychology*, April, 1924.

† It is also possible that formula (94) does not apply when $r = .987$, and $N = 116$.

$$P.E._{M_1 - M_2} = \sqrt{(P.E._{M_1})^2 + (P.E._{M_2})^2 - 2r_{12}P.E._{M_1}P.E._{M_2}}. \quad (100)$$

{Probable error of difference between means where correlated}

This formula may be illustrated by a comparison of the length of the left forearm for 1063 English males and their adult sons.* The results found were

$$M_S = 18.52'' \pm 0.021'', \text{ and } M_F = 18.31'' \pm 0.019'',$$

while r_{FS} was equal to .421, the size of forearm in father and son showing considerable correlation. Substituting in formula (100), we find that

$$P.E._{M_1 - M_2} = \sqrt{(.021)^2 + (.019)^2 - 2(.421)(.021)(.019)} = .022.$$

The difference may then be written $0.21'' \pm .022$. Since this is about nine times its probable error, there is no doubt that the sons of the professional English class were substantially differentiated from their fathers by a slightly longer forearm.

6. The Probable Errors of Observed and Percentage Frequencies

In comparing the frequencies between two groups it is often convenient to reduce them to percentages as in the table on page 244 taken from columns 1 and 2 of Table 55.

If f denotes an observed frequency, its probable error is given by the formula

$$P.E._f = .6745\sqrt{f\left(1 - \frac{f}{N}\right)},\dagger \left\{\begin{array}{c}\text{Probable error of an}\\ \text{observed frequency}\end{array}\right\} \quad (101)$$

while for a percentage frequency $f_p = \frac{100f}{N}$, we have

$$P.E._{f_p} = .6745\sqrt{\frac{f_p(100 - f_p)}{N}}.\ddagger \left\{\begin{array}{c}\text{Probable error of a per-}\\ \text{centage frequency}\end{array}\right\} \quad (102)$$

* *Biometrika*, Vol. II, p. 370.

† This formula may be derived from equation (105) by setting $p = \frac{f}{N}$ and $q = \left(1 - \frac{f}{N}\right)$. For a complete and excellent proof see Jones, op. cit., p. 151.

‡ Derived from formula (106) by finding the $P.E.$ of 100 p, or f_p.

TABLE 56. FREQUENCY PERCENTAGES OF I. Q.'S FOR GRADE AND HIGH SCHOOL WHITE CHILDREN

I.Q.	FREQUENCY PERCENTAGES	
	Grade Schools	High Schools
150–160	0.1	—
140–150	0.6	—
130–140	1.7	0.3
120–130	4.7	3.1
110–120	11.3	17.5
100–110	22.5	39.3
90–100	26.5	28.5
80–90	17.9	10.8
70–80	11.2	0.5
60–70	2.6	—
50–60	0.6	—
40–50	0.2	—
30–40	0.1	—
Total	100.0	100.0

Applying formula (102) to the percentage frequencies in the interval 100 to 110, we find

$$39.3 \pm .6745\sqrt{\frac{39.3(60.7)}{389}}, \text{ or } 39.3 \pm 1.67,$$

and

$$22.5 \pm .6745\sqrt{\frac{22.5(77.5)}{1560}}, \text{ or } 22.5 \pm 0.71,$$

$$P.E.(diff.) = \sqrt{(1.67)^2 + (0.71)^2} = 1.81.$$

The difference $39.3 - 22.5$ may therefore be written 16.8 ± 1.81. We may conclude that a significantly higher percentage of high-school pupils is found in the group with I. Q.'s between 100 and 110.

Formula (101) is often useful in comparing observed with theoretical frequencies. Thus in Fig. 55 the area under the normal curve from 80 to 90 is larger than that given by the column of the histogram. In order to find the area under the curve it is necessary to express the class limits as deviates from the mean and enter a table of areas such as Holzinger's Table XI. The arithmetic will then be as follows:

$$\frac{x_1}{\sigma} = \frac{80 - 89.28}{16.61} = -0.559, \qquad \frac{x_2}{\sigma} = \frac{90 - 89.28}{16.61} = +0.043,$$
$$\tfrac{1}{2}\alpha = .2120,^* \qquad \tfrac{1}{2}\alpha = .0172.$$

Therefore, the normal frequency is $4834(.2120 + .0172)$, or 1108.

From formula (101) the probable error of the observed frequency 1059 is $.6745\sqrt{1059 \times .7809} = 19$. The difference $1108 - 1059 = 49 \pm 19$ might therefore be attributable to the fluctuations of sampling.

7. The Chi-Square Test

In the case of a whole frequency distribution such as for the 4834 I. Q.'s, a comparison of the observed and theoretical frequencies may be made by Pearson's Chi-Square Test. Any such distribution is to be regarded as a sample from a much larger group. The problem is then to determine whether or not the fitted curve is a sufficiently good description of the observed data within the fluctuations of sampling.

The test is made by obtaining all the differences between observed and theoretical frequencies, substituting the result in a formula, and determining by a table the probability that random sampling would give as bad a fit or worse.

If the observed frequencies are denoted by

$$f'_1, f'_2, f'_3, \cdots f'_n$$

and the corresponding theoretical frequencies by

$$f_1, f_2, f_3 \cdots f_n,$$

the value for χ^2 may be written

$$\chi^2 = \sum_{t=1}^{t=n} \left\{ \frac{(f'_t - f_t)^2}{f_t} \right\} \cdot \left\{ \begin{matrix} \text{Chi-square} \\ \text{function} \end{matrix} \right\} \quad (103)$$

* These values have been obtained from Table XI by linear interpolation, that is, when $\frac{x}{\sigma} = .55$, $\frac{1}{2}\alpha = .2088$ and when $\frac{x}{\sigma} = .56$, $\frac{1}{2}\alpha = .2123$. The value of $\frac{1}{2}\alpha$ for $\frac{x}{\sigma} = .559$ is therefore .9 of the difference $.0035 + .2088$, or .2120.

Before taking up the probability test we shall next work out χ^2 for the distribution of 4834 I.Q.'s fitted by a normal curve.

In determining the values of f_t it will be found convenient to obtain the fractional area from the mean to the limits of the various groups, and then subtract these values successively and multiply by N to give the theoretical frequencies comparable with f'_t. The complete arithmetic for the values of f_t is shown in the following table:

TABLE 57. SHOWING THE CALCULATION OF NORMAL FREQUENCIES

GROUP LIMITS X_t	$X_t - M$	$\frac{X_t - M}{\sigma}$ ($\sigma = 16.61$)	AREA FROM M*	AREA IN GROUPS X_{t-1} TO X_t	f_t = LAST VALUES × 4834
160	+ 70.72	+ 4.26	.5000	.0001	0.5
150	+ 60.72	+ 3.66	.4999	.0010	4.8
140	+ 50.72	+ 3.05	.4989	.0060	29.0
130	+ 40.72	+ 2.45	.4929	.0251	121.3
120	+ 30.72	+ 1.85	.4678	.0734	354.8
110	+ 20.72	+ 1.25	.3944	.1522	735.7
100	+ 10.72	+ 0.65	.2422	.2262	1093.5
90	+ 0.72	+ 0.04	.0160	.2283†	1103.6
80	− 9.28	− 0.56	.2123	.1647	796.2
70	− 19.28	− 1.16	.3770	.0838	405.1
60	− 29.28	− 1.76	.4608	.0301	145.5
50	− 39.28	− 2.36	.4909	.0076	36.7
40	− 49.28	− 2.97	.4985	.0013	6.3
30	− 59.28	− 3.57	.4998	.0002‡	1.0
Total				1.0000	4834.0

The only point where any difficulty is likely to arise is in passing over the group containing the mean. Here the frequencies must be added to obtain the frequency of the group. A rough diagram of the normal curve will clarify the whole procedure.

In working out χ^2 Professor Pearson recommends the consolidation of the small frequencies in the end groups. For the top interval we shall therefore add 0.5 and 4.8. The excess of 1.0 below 30 may also be added to the lowest group to give 7.3. Table 58 then shows the remainder of the calculation.

* By Holzinger's Table XI. † (.2123 + .0160 = .2283.) ‡ .0002 is below 30.

The number of frequency groups is denoted by n'. Entering Pearson's Table XII with $n' = 12$ and $\chi^2 = 46.6$, we find $P = .00001$. The interpretation of this result is that once in 100,000 trials we should get, in random sampling, a fit as bad or worse than that which would be obtained if the real distribution were represented by the normal curve fitted above. The actual fit is therefore a very bad one. Unless the value of P be .2 or more, the fit cannot be regarded as good and other curves should be tried.

The importance of the χ^2 test arises from the fact that it furnishes a rigorous method for determining goodness of fit.

TABLE 58. SHOWING THE CALCULATION OF χ^2

CLASS	OBSERVED FREQUENCY $f't$	THEORETICAL FREQUENCY ft	$f't - ft$	$(f't - ft)^2$	$\frac{(f't - ft)^2}{ft}$
140–160	14	5.3	+ 8.7	75.69	14.3
130–140	36	29.0	+ 7.0	49.00	1.7
120–130	103	121.3	− 18.3	334.89	2.8
110–120	318	354.8	− 36.8	1354.24	3.8
100–110	799	735.7	+ 63.3	4006.89	5.4
90–100	1074	1093.5	− 19.5	380.25	0.3
80–90	1059	1103.6	− 44.6	1989.16	1.8
70–80	868	796.2	+ 71.8	5155.24	6.5
60–70	366	405.1	− 39.1	1528.81	3.8
50–60	163	145.5	+ 17.5	306.25	2.1
40–50	25	36.7	− 11.7	136.89	3.7
30–40	9	7.3	+ 1.7	2.89	0.4
Total	4834	4834.0	00.0		46.6

Mere inspection of the data is of no value except to suggest the theoretical form of the curve to be fitted. When this has been selected by guess (or by the method of Chapter XVI) the fit should be tested by a procedure similar to that shown above. Other uses of the χ^2 function will be given in Chapter XIV.

A very much abbreviated table for the values of P is given in Table 59 on page 248 for use when χ^2 and n' are not large. This table has been taken from Pearson's Table XII, the computation of which was done by Mr. W. P. Elderton.

TABLE 59. VALUES OF P FOR TESTING GOODNESS OF FIT

χ^2	n'								
	7	8	9	10	11	12	13	14	15
1	.986	.995	.998	.999	1.000	1.000	1.000	1.000	1.000
2	.920	.960	.981	.991	.996	.998	.999	1.000	1.000
3	.809	.885	.934	.964	.981	.991	.996	.998	.999
4	.677	.780	.857	.911	.947	.970	.983	.991	.995
5	.544	.660	.758	.834	.891	.931	.958	.975	.986
6	.423	.540	.647	.740	.815	.873	.916	.946	.966
7	.321	.429	.537	.637	.725	.799	.858	.902	.935
8	.238	.333	.433	.534	.629	.713	.785	.844	.889
9	.174	.253	.342	.437	.532	.622	.703	.773	.831
10	.125	.189	.265	.350	.440	.530	.616	.694	.762
11	.088	.139	.202	.276	.358	.443	.529	.611	.686
12	.062	.101	.151	.213	.285	.363	.446	.528	.606
13	.043	.072	.112	.163	.224	.293	.369	.448	.527
14	.030	.051	.082	.122	.173	.233	.301	.374	.450
15	.020	.036	.059	.091	.132	.182	.241	.307	.378

8. THE PROBABLE ERROR OF AN OBSERVED PROPORTION

It has already been shown in Chapter XI that the mean and standard deviation of the point binomial $(q + p)^n$ are given by np and $\sqrt{npq}$, respectively. In the case where we are dealing with K samples of n events each, the binomial becomes $K(q + p)^n$ for which the mean and standard deviation are the same as before.

Now if the proportion of successes instead of the actual number is recorded, it will be necessary to take one nth of the number in each sample. The mean proportion of successes will then approach p and the standard deviation will be given by

$$\sigma_p = \sqrt{\frac{pq}{n}}. \qquad \left\{\begin{matrix}\text{Standard error}\\ \text{of a proportion}\end{matrix}\right\} \qquad (104)$$

The equations for probable errors of the mean number and of the proportion of successes in a sample are therefore

$$P.E._{np} = .6745\sqrt{npq}, \qquad \left\{\begin{matrix}\text{Probable errors of the}\\ \text{mean and of the pro-}\\ \text{portion of successes}\end{matrix}\right\} \qquad (105)$$

and

$$P.E._p = .6745\sqrt{\frac{pq}{n}}, \qquad (106)$$

respectively.

This last formula may be illustrated by the use of some data taken from the 1920–1921 Register of The University of Chicago. The total number of students for that year may be tabulated in the following form:

	Men	Women	Total
Graduate schools (group 1)	1,433	1,246	2,679(n_1)
Undergraduate schools (group 2)	3,938	4,768	8,706(n_2)
Total	5,371	6,014	11,385

The problem is to determine whether or not the proportion of men in the graduate schools is significantly larger than in the undergraduate schools. In this case a "success" is given by the registration of a man and a "failure" by the registration of a woman, while the total for each is the size of the sample, n.

The observed proportion of men in the graduate schools is $\frac{1433}{2679} = .535 = p_1$, while the proportion of men undergraduates is $\frac{3938}{8706} = .452 = p_2$. It is also evident that $q_1 = .465$, $n_1 = 2679$, $q_2 = .548$, and $n_2 = 8706$. From formula (106) we therefore have

$$p_1 = .535 \pm .6745\sqrt{\frac{(.535)(.465)}{2679}} = .535 \pm .0065,$$

and

$$p_2 = .452 \pm .6745\sqrt{\frac{(.452)(.548)}{8706}} = .452 \pm .0036.$$

The difference between the two proportions may therefore be written

$$p_1 - p_2 = .083 \pm \sqrt{(.0065)^2 + (.0036)^2} = .083 \pm .0074.$$

Assuming that the observed proportions are typical of other years, or that the above data furnish random samples, we may conclude that the graduate schools enroll a significantly larger proportion of men graduates. It should be noted, however, that the conditions brought about by the war might invalidate such assumptions. The safest procedure, therefore, would be to calculate the differences for a number of years.

Another method of approach to the above problem is to determine whether or not the difference between the two proportions could have arisen merely from the fluctuations in sampling in case the two groups are regarded as samples from the same or very similar populations.

The proportion of men in both schools is given by $p_0 = \frac{5371}{11385} = .472$, with $q_0 = .528$. The equations for the probable errors of the proportions in the two samples will then be

$$P.E._{p_1} = .6745\sqrt{\frac{p_0q_0}{n_1}}, \quad \left\{\begin{array}{c}\text{Probable errors of proportions of successes, based on both groups}\end{array}\right\} \tag{107a}$$

and

$$P.E._{p_2} = .6745\sqrt{\frac{p_0q_0}{n_2}}. \tag{107b}$$

Applying these formulas to the above data, we have

$$P.E._{p_1} = .6745\sqrt{\frac{(.472)(.528)}{2679}} = .0065,$$

and

$$P.E._{p_2} = .6745\sqrt{\frac{(.472)(.528)}{8706}} = .0036,$$

agreeing to four places with the results found by formula (106). The difference test, of course, gives $p_1 - p_2 = .083 \pm .0074$ as before, and we may therefore safely conclude that random sampling could not have accounted for the difference between the observed proportions. The difference between the values given by formulas (106) and (107) is chiefly a theoretical one, for they do not differ largely unless p_1 and p_2 differ largely.

9. Response Error Formulas

A number of formulas for dealing with the response error described in Chapter V will next be obtained. The notation to be employed may be given as follows:

z_1 and z_I = standard scores on two forms of X_1,
z_2 and z_{II} = standard scores on two forms of X_2,

r_{1I} and r_{2II} = reliability coefficients of X_1 and X_2, respectively,
e_1 and e_I = response errors in X_1 by two forms,
e_2 and e_{II} = response errors in X_2 by two forms, and
s and t = average or "true" scores of an individual on X_1 and X_2, respectively.

We may therefore write

$$\left.\begin{aligned} z_1 &= s + e_1, \\ z_I &= s + e_I, \\ z_2 &= t + e_2, \\ z_{II} &= t + e_{II}. \end{aligned}\right\} \quad \left\{\begin{array}{c}\text{Standard scores in terms of "true" scores and response error}\end{array}\right\} \tag{108}$$

It will be assumed in the following proofs that the response errors, e, are not correlated with each other nor with the true scores, s and t. While this assumption is a reasonable one, it is not necessarily valid and the resulting formulas should be used with caution pending a verification of these assumptions.

If two forms of a test are given we may write

$$z_1 - z_I = e_1 - e_I.$$

Squaring, summing for a group of individuals, and dividing by N, there results

$$\frac{\Sigma z_1^2}{N} - 2\frac{\Sigma z_1 z_I}{N} + \frac{\Sigma z_I^2}{N} = \frac{\Sigma e_1^2}{N} - 2\frac{\Sigma e_1 e_I}{N} + \frac{\Sigma e_I^2}{N},$$

or

$$2 - 2\,r_{1I} = 2\,\sigma_{e_1}^2,$$

since $\sigma_z = 1$, $\sigma_{e_1} = \sigma_{e_I}$, and $r_{e_1 e_I} = 0$.

The required response error formula therefore becomes

$$\sigma_{e_1} = \sqrt{1 - r_{1I}}, \qquad \left\{\begin{array}{c}\text{Standard error of response using } z \text{ scores}\end{array}\right\} \tag{109}$$

or, if the original scores X are used,

$$\sigma_{e_1} = \sigma_{x_1}\sqrt{1 - r_{1I}}. \quad \left\{\begin{array}{c}\text{Standard error of response using } X \text{ scores}\end{array}\right\} \tag{110}$$

These formulas give the standard deviation of the N errors e_1 for a group and thus furnish an approximation to the standard deviation of many similar errors for a single individual. They

measure, therefore, the standard error of response within an average individual of the group. In case the probable error is used as a unit we have

$$P.E._{e_1}\text{ (of individual } X_1) = .6745\,\sigma_{x_1}\sqrt{1 - r_{1I}}. \qquad (111)$$

{Probable error of response for X_1}

As an example, consider a test with a reliability coefficient of .64 and standard deviation of 5. Substituting these values in (111), we find that $P.E._{e_1} = .6745 \times 5 \times .6 = 2.02$. A pupil's score, such as 31, may therefore be written 31 ± 2 with the interpretation that it is an even chance that his true score, assuming that there is no practice effect, will lie anywhere between 29 and 33. To facilitate calculations of this sort, values of $\sqrt{1 - r}$ have been prepared and tabled in Holzinger's Tables for Students, No. VIII.

Returning to equations (108) and (109), we may next find the response error of the difference $z_1 - z_2$ between two tests which may be quite dissimilar. The quantity required is $\sigma_{(e_1 - e_2)}$ or $\sigma_{(e_I - e_{II})}$. Since the errors e are all uncorrelated

$$\sigma^2_{(e_1 - e_2)} = \frac{\Sigma e_1^2}{N} + \frac{\Sigma e_2^2}{N} - \frac{2\,\Sigma e_1 e_2}{N} = \sigma^2_{e_1} + \sigma^2_{e_2}.$$

From (109) the error $\sigma_{e_1} = \sqrt{1 - r_{1I}}$, and $\sigma_{e_2} = \sqrt{1 - r_{2II}}$ by similar proof. Substituting these values in the above equation, and taking the square root of both members, we find that

$$\sigma_{(e_1 - e_2)} = \sqrt{2 - r_{1I} - r_{2II}},^* \left\{\begin{matrix}\text{Standard error of}\\ \text{response for } z_1 - z_2\end{matrix}\right\} \qquad (112)$$

or

$$P.E.\text{ (of individual } z_1 - z_2) = .6745\sqrt{2 - r_{1I} - r_{2II}}. \qquad (113)$$

To illustrate this formula consider two tests, say in arithmetic and spelling, given in a school grade, and assume the reliability of both tests to be .5. The difference between two standard scores, say 2.6 and 1.4, for a given pupil may therefore

* This formula was first derived by T. L. Kelley in *Journal of Educational Research*, September, 1923. A note on his proof is given by the writer in the January, 1925, issue of the same journal.

be written $2.6 - 1.4 = 1.2 \pm .6745$. Since this difference is approximately twice its probable error the chances are about four to one that the true difference lies between zero and 2.4.

An observed standard deviation will be larger than the true standard deviation because of the effect of response errors. This may be shown by writing

$$x_1 = s + e_1,$$

whence $$\sigma_{x_1}^2 = \sigma_s^2 + \sigma_{e_1}^2.$$

From equation (110) $$\sigma_{e_1}^2 = \sigma_{x_1}^2 - \sigma_{x_1}^2 r_{1I},$$

so that $$\sigma_s = \sigma_{x_1}\sqrt{r_{1I}}. \quad \left\{\begin{matrix}\text{Relation between true and}\\ \text{observed standard errors}\end{matrix}\right\} \qquad (114)$$

It is therefore apparent that only for a perfectly reliable test will the observed and true standard deviations be equal.

Professor Spearman* has given a number of formulas for correcting correlation coefficients for response error, or "attenuation" as he calls it. One of the simplest of these may be worked out as follows:

The correlation between "true" scores is $r_{st} = \dfrac{\Sigma st}{N\sigma_s\sigma_t}$. But $\Sigma st = \Sigma z_1 z_2$, from equation (108), while $\sigma_s = \sqrt{r_{1I}}$ and $\sigma_t = \sqrt{r_{2II}}$.

Therefore, $$r_{st} = \frac{r_{12}}{\sqrt{r_{1I}r_{2II}}}, \quad \left\{\begin{matrix}\text{Spearman's correction}\\ \text{for attenuation}\end{matrix}\right\} \qquad (115)$$

where r_{12} is the observed correlation.†

As an example of the use of formula (115), if an observed correlation is .6 and the reliability coefficients of X_1 and X_2 are both .8, the "true" correlation, with response error eliminated, will be .75.

If σ and Σ denote the standard deviations on a test for two groups, and r_{1I} and R_{1I} the respective reliability coefficients, it is evident from formula (110) that

$$\sigma_e = \sigma\sqrt{1 - r_{1I}} \quad \text{and} \quad \sigma_E = \Sigma\sqrt{1 - R_{1I}}.$$

* C. Spearman, "Demonstration of Formulæ for True Measurement of Correlation," *American Journal of Psychology*, Vol. XVIII (1907), p. 161, and "Correlation from Faulty Data," *British Journal of Psychology*, Vol. III (1910), p. 271.

† For other correction formulas see Yule, Introduction to Statistics, p. 213.

Assuming with Professor Kelley* that $\sigma_e = \sigma_E$, or that the test is "equally effective" for both groups, we find that

$$\frac{\sigma}{\Sigma} = \frac{\sqrt{1 - R_{1I}}}{\sqrt{1 - r_{1I}}}, \qquad \left\{\begin{array}{c}\text{Kelley's formula for}\\ \text{adjusting reliability}\\ \text{coefficients}\end{array}\right\} \tag{116a}$$

or

$$r_{1I} = \frac{\sigma^2 - \Sigma^2(1 - R_{1I})}{\sigma^2}. \tag{116b}$$

This formula has been used to adjust correlations for different ranges as illustrated by the following examples. If the reliability of a test is given by $R_{1I} = .5$ for a range with $\Sigma = 5$, what will the reliability be for a range with standard deviation of 10? From (116b) we find $r_{1I} = .875$, which shows the effect of "range of talent" upon the reliability coefficient. It should be noted, however, that for very small values formula (116) gives results of doubtful significance. Thus when $\Sigma = 5$, $\sigma = 10$, and $R_{1I} = .01$, we find $r_{1I} = .75$. That a test which is practically worthless on one range should be quite reliable on range with twice as great variability is contrary to all experience with such measures.

A general criticism of all the above formulas is that the assumption of uncorrelated response errors does not appear to be justified.† Such negative evidence, however, is not sufficient at present to warrant the entire abandonment of the formulas, and they are offered here for tentative use until further evidence in proof is available.

EXERCISES

1. Find the probable errors of the frequencies at I.Q. 80–90 given in the columns of Table 55. (10.2, 4.1, 12.1, 9.4, 3.7, 19.4. *Ans.*)

2. Determine the probable errors of the following correlation coefficients: $r = .162$ $(N = 87)$, $r = .083$ $(N = 640)$, $r = .204$ $(N = 49)$, $r = -.137$ $(N = 210)$, $r = .083$ $(N = 40)$. Use Holzinger's Table X. (.070, .026, .092, .046, .106. *Ans.*)

* Kelley, Statistical Method, p. 222. See also Chapter IX of the present text.

† William Brown and Godfrey H. Thomson, in "Essentials of Mental Measurement" (Cambridge University Press, England, 1921), show correlation between such errors.

3. Test the significance of the differences between the means and standard deviations given in Table 55. Use Table 54.

4. The following data were obtained from four groups:

$M_1 = 104$	$\sigma_1 = 10.0$	$N_1 = 110$
$M_2 = 101$	$\sigma_2 = 11.0$	$N_2 = 97$
$M_3 = 102$	$\sigma_3 = 9.6$	$N_3 = 92$
$M_4 = 103$	$\sigma_4 = 8.5$	$N_4 = 106$

Find the probabilities that M_1 will be larger than M_2, M_3, and M_4, respectively, on the next sampling. (.98, .92, .79. *Ans.*)

HINT. Use Table 54.

5. In a six-month period, 454 deaths from automobiles were reported in New York and 260 in Chicago. The populations of the two cities were 5,600,000 and 2,700,000, respectively. Are "Gotham's streets safer for the pedestrian than Chicago's," as reported by a certain newspaper? (Difference in death rates is three times its *P.E.*)

6. The following data were taken from the President's Report of The University of Chicago, 1923–1924.

	MEN	WOMEN	TOTAL
Graduate schools	2,083	1,634	3,717
Undergraduate schools	4,215	5,425	9,640
Total	6,298	7,059	13,357

Find the proportion of men in the graduate and in the undergraduate schools, and test the significance of the difference found.

($p_1 - p_2 = .123 \pm .0065$. *Ans.*)

7. Fit, with a normal curve, the distribution of the Terman scores given in Exercise 3 of Chapter II, and apply the χ^2 test.

($P \doteq .6$. *Ans.*)

8. Apply the χ^2 test to the distributions of I.Q.'s fitted in Exercise 9 of Chapter XII.

CHAPTER XIV

FURTHER METHODS OF CORRELATION FOR TWO CHARACTERS

1. Introductory

The correlation methods discussed thus far have been those which are applied to quantitative series or to traits which are measurable on a numerical scale. In case the series are qualitative or unordered, in the sense used in the second chapter, other methods for measuring the association become necessary. The present chapter will therefore be concerned with the treatment of such series by suitable methods.

In order to illustrate the combinations of series that may arise, we may begin by listing some of the possibilities with short supposititious examples. The table below illustrates the case of an association for quantitative and qualitative series, intelligence being measured on a numerical scale and school work rated in verbal categories in orderly progression.

Table 60. Illustrating Associated Quantitative and Qualitative Series

School Work	I.Q.				
	80	90	100	110	120
Good	3	12	14	11	Qualitative
Medium	4	15	17	2	
Poor	7	3	12	—	
	Quantitative				

Table 61 on page 257 shows the association between two qualitative series, both characteristics being verbally indexed.

TABLE 61. ILLUSTRATING ASSOCIATED QUALITATIVE SERIES

SCHOOL WORK	BEHAVIOR				
	Bad	Troublesome	Good	Excellent	
Good	3	9	12	14	QUALITATIVE
Medium	4	10	16	2	
Poor	10	2	7	—	
	QUALITATIVE				

In both tables there appears to be some association between the traits, but it cannot be adequately measured by the product-moment correlation in the form used in Chapter IX, because of the lack of numerical indexes for the categories.

An example of association for quantitative and unordered series is next given in Table 62, the characteristics being the intelligence of children and the occupation of their fathers.

TABLE 62. ILLUSTRATING ASSOCIATED QUANTITATIVE AND UNORDERED SERIES

OCCUPATION OF FATHER	I. Q. OF CHILD				
	80	90	100	110	120
Teacher	7	11	12	10	UNORDERED
Doctor	3	9	14	8	
Lawyer	3	6	9	12	
Writer	3	4	7	11	
	QUANTITATIVE				

The relationship in this case cannot be observed very readily, because the arrangement of the occupation categories is a matter of indifference. A quite different method of measuring association will therefore be required for such a problem.

A complete list of the combinations of series which may arise is given as follows:

a. Quantitative with quantitative
b. Quantitative with qualitative
c. Quantitative with unordered
d. Qualitative with qualitative
e. Qualitative with unordered
f. Unordered with unordered

While some of these occur only rarely in statistical work, it is nevertheless desirable to have suitable methods for dealing with each type of association. The methods, however, are by no means restricted to one type of problem, and consequently the choice often becomes a difficult matter. In the present discussion we shall select a few of the outstanding methods available and apply them to problems with suggestions as to the appropriate method to employ whenever possible.

2. Another Formula for the Product-Moment Method

Before taking up the correlation of qualitative series we shall first introduce a modification of the product-moment formula convenient for dealing with such data. The method was presented by Professor Pearson in one of his lectures at the University of London.

Using the notation of Chapter IX, the product-moment formula may be written

$$r = \frac{\left[\Sigma f_{xy} d_x d_y - \frac{(\Sigma f_x d_x)(\Sigma f_y d_y)}{N}\right] hk}{N\sigma_x \sigma_y}, \tag{117}$$

where the product hk occurs because the numerator is expressed in class intervals. If $\overline{Y}_x$ denotes the mean of a column and M_y the mean of the whole table, it is also evident that

$$\overline{Y}_x - M_y = \left(\frac{\Sigma' f_{xy} d_y}{f_x} - \frac{\Sigma f_y d_y}{N}\right) k.$$

Multiplying both members of this equation by hd_x, summing over the whole table, and noting that f_x is merely a symbol of operation, we have

$$\Sigma f_x d_x(\overline{Y}_x - M_y)h = \left[\Sigma f_{xy} d_x d_y - \frac{(\Sigma f_x d_x)(\Sigma f_y d_y)}{N}\right]hk.$$

Substituting this result in formula (117), we then obtain

$$r = \frac{\Sigma f_x d_x(\overline{Y}_x - M_y)h}{N\sigma_x\sigma_y}, \tag{118a}$$

and, similarly,

$$r = \frac{\Sigma f_y d_y(\overline{X}_y - M_x)k}{N\sigma_x\sigma_y}. \tag{118b}$$

{ Pearson's formulas for the correlation coefficient based on the means of the arrays }

The above method is very convenient when the means of the arrays are known, for it is then not necessary to calculate the quantity $\Sigma f_{xy} d_x d_y$ from the individual cells. It should be noted that the variables d_x and d_y may be taken from any origins whatsoever, and it may seem a little curious at first that the values of formulas (118 a) and (118 b) remain unchanged when the origins are shifted and all quantities except d_x and d_y are fixed throughout in these formulas.

TABLE 63. ILLUSTRATING THE CALCULATION OF THE CORRELATION COEFFICIENT BY FORMULA (118 a)

X	$\overline{Y}_x$	$\overline{Y}_x - M_y$	f_x	d_x	$(\overline{Y}_x - M_y)f_x d_x h$
184.5	72.25	+18.50	1	5	+925.00
174.5	52.25	−1.50	1	4	−60.00
164.5	64.75	+11.00	4	3	+1320.00
154.5	60.89	+7.14	11	2	+1570.80
144.5	57.25	+3.50	9	1	+315.00
134.5	52.70	−1.05	11	0	0.00
124.5	45.25	−8.50	5	−1	+425.00
114.5	42.25	−11.50	4	−2	+920.00
104.5	37.25	−16.50	2	−3	+990.00
94.5	37.25	−16.50	1	−4	+660.00
84.5	32.25	−21.50	1	−5	+1075.00
			50		8140.80

$M_y = 53.75$ $\quad \sigma_x = 19.92$ $\quad \sigma_y = 10.50$ $\quad N\sigma_x\sigma_y = 10{,}458$

$$\therefore\ r = \frac{8140.8}{10458} = +.778$$

In order to illustrate the application of this method to quantitative data, the correlation problem shown in Table 30 of Chapter IX has been worked out on page 259, using formula (118a). The means of the columns $\overline{Y}_x$ were calculated as for any distribution, and the values for d_x were taken from the arbitrary origin 134.5. A check on the numerator of (118a) may be made by shifting to another origin and recalculating the sum of all the products. The proof of this check is left as an exercise.

3. The Product-Moment Method for Qualitative Series

A qualitative series may be converted into a quantitative one by representing the data on a normal scale as shown in section 7 of Chapter XII. The various groups will then be designated by numbers instead of by verbal description, and the product-moment method may then be applied for measuring the amount of correlation.

The following table represents the correlation between the score on a physics test and the rating of the teachers for 245 high-school pupils. The combination is, therefore, a quantitative series with a qualitative one, and the latter will need to be converted to a normal scale.

Table 64. Data from a Physics Test and Teacher Rating

Test Score	Teacher Rating				Total
	Poor	Fair	Good	Excellent	
70–80	—	—	2	2	4
60–70	1	6	12	18	37
50–60	11	15	24	18	68
40–50	19	26	23	16	84
30–40	10	17	9	4	40
20–30	6	2	1	1	10
10–20	—	1	—	—	1
0–10	1	—	—	—	1
Total	48	67	71	59	245
Per cent	19.6	27.3	29.0	24.1	100.0

The ordinates bounding the various pieces under the normal curve are most readily found by entering a table such as Holzinger's Table XII with the cumulative frequencies .196, .469, and .759, each less .5, or with the values $-.304$, $-.031$, and .259. The three ordinates resulting are .2766, .3977, and .3116, respectively, as illustrated in Fig. 66.

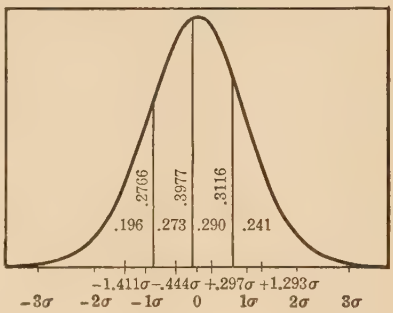


FIG. 66. Illustrating the means of the four rating categories when the series is represented on a normal scale

The means of the various pieces may now be worked out by formula (86) of Chapter XII; for example,

$$\frac{\bar{x}_p}{\sigma_x} = \frac{0 - .2766}{.196} = -1.411,$$

where $\frac{\bar{x}_p}{\sigma_x}$ is the mean of the "poor" category. For the other three means, we obtain $-.444$, .297, and 1.293. These numbers are to be regarded as class values in the subsequent calculations.

Since $M_x = 0$ for a normal distribution, the required formula may be obtained from (118b) in the form

$$r = \frac{\Sigma f_y d_y \left(\frac{\bar{x}_y}{\sigma_x}\right) k}{N\sigma_y}, \quad \left\{\begin{array}{l}\text{Correlation coefficient}\\ \text{adapted for use with}\\ \text{data on a normal scale}\end{array}\right\} \qquad (119)$$

where $\frac{\bar{x}_y}{\sigma_x}$ denotes the mean of a row measured from the mean of the table. The values for $\frac{\bar{x}_y}{\sigma_x}$ are obtained by multiplying the frequencies in each row by the class values just obtained, and dividing by the total in the row. Thus for the top and next row,

$$\frac{\bar{x}_{75}}{\sigma_x} = \frac{2 \times .297 + 2 \times 1.293}{4} = +.795,$$

$$\frac{\bar{x}_{65}}{\sigma_x} = \frac{1(-1.411) + 6(-.444) + 12(.297) + 18(1.293)}{37} = +.615,$$

and, similarly, $\frac{\bar{x}_{55}}{\sigma_x} = +.121$, $\frac{\bar{x}_{15}}{\sigma_x} = -.129$, $\frac{\bar{x}_{35}}{\sigma_x} = -.345$, $\frac{\bar{x}_{25}}{\sigma_x} = -.776$, $\frac{\bar{x}_{15}}{\sigma_x} = -.444$, and $\frac{\bar{x}_{5}}{\sigma_x} = -1.411$.

An arrangement of the computation for the product sum and σ_y is shown below. The work is best done with a machine.

TABLE 65. ILLUSTRATING THE CALCULATION OF THE CORRELATION COEFFICIENT FOR THE DATA IN TABLE 64

f_y	d_y	$f_y d_y$	$\frac{\bar{x}_y}{\sigma_x}$	$f_y d_y \frac{\bar{x}_y}{\sigma_x}$	$f_y d_y^2$
4	3	12	.795	9.540	36
37	2	74	.615	45.510	148
68	1	68	.121	8.228	68
84	0	0	−.129	0.000	0
40	−1	−40	−.345	13.800	40
10	−2	−20	−.776	15.520	40
1	−3	−3	−.444	1.332	9
1	−4	−4	−1.411	5.644	16
245		+87		99.574	357

$\sigma_y = 11.54 \quad N\sigma_y = 2827.3$

$\therefore r = \frac{99.574 \times 10}{2827.3} = .352$

By plotting the means of the rows $\frac{\bar{x}_y}{\sigma_x}$ as shown in Fig. 67, a graphical representation of the regression is given. It will be noted that the points fall fairly closely along a straight line, so that the regression is probably to be regarded as linear. The equation of the regression line through the mean of the table is $\frac{\bar{x}}{\sigma_x} = \frac{r}{\sigma_y} y$, or $\frac{\bar{x}}{\sigma_x} = .0305\, y$. Since $M_y = 48.55$, two points for plotting are given by substituting $y = \pm 30$ in the above equation or,

$$\begin{cases} \frac{\bar{x}}{\sigma_x} = .915 \\ Y = 78.55 \end{cases} \quad \text{and} \quad \begin{cases} \frac{\bar{x}}{\sigma_x} = -.915 \\ Y = 18.55 \end{cases}$$

When both series are qualitative, the above method may be applied to the two scales; but, since certain corrections are

sometimes desirable, another procedure will be shown. In calculating the correlation coefficient and other measures of association an error is introduced by grouping the material in broad categories. Professor Pearson* has devised several formulas for correcting this error, one of which may be written in the form

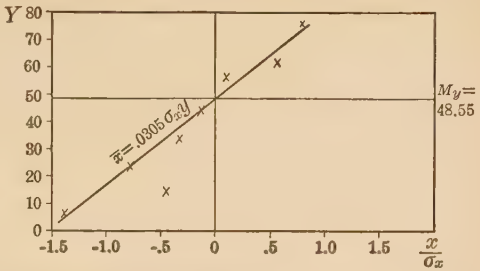


FIG. 67. Regression line for the physics data

$$ {}_c r_{xy} = \frac{\Sigma f_{xy} \dfrac{N}{f_x f_y} (z_s - z_{s+1})(z'_s - z'_{s+1})}{\left[\Sigma \dfrac{N}{f_x} (z_s - z_{s+1})^2\right]\left[\Sigma \dfrac{N}{f_y} (z'_s - z'_{s+1})^2\right]}, \qquad (120) $$

{Pearson's corrective formula for broad grouping assuming normal distributions of the variates}

where the z's are ordinates bounding the various pieces under the normal curve and the unprimed and primed values refer to X and Y, respectively. The use of this formula will next be illustrated by a problem which has been taken from Professor Pearson's paper cited in footnote below.

TABLE 66. PEARSON'S DATA ON INTELLIGENCE AND QUALITY OF CLOTHING

QUALITY OF CLOTHING	INTELLIGENCE RATING						TOTAL
	B	*C*	*D*	*E*	*F*	*G*	
I	33	48	113	209	194	39	636
II	41	100	202	255	138	15	751
III	39	58	70	61	33	4	265
IV and V	17	13	22	10	10	1	73
Total	130	219	407	535	375	59	1725

* Karl Pearson, "On the Measurement of the Influence of Broad Categories upon Correlation," *Biometrika*, Vol. IX, p. 119.

TABLE 67. SHOWING THE CALCULATION FOR FORMULA (120)

							f_y	f_y/N	z'_s	$z'_s - z'_{s+1}$
$f_{xy}\rightarrow$	33	48	113	209	194	39	636	.369	.3772	+.3772
$f_x f_y\rightarrow$	82,680	139,284	258,852	340,260	238,500	37,524				
$\frac{Nf_{xy}}{f_x f_y}\rightarrow$	.6885	.5945	.7530	1.0596	1.4031	1.7929				
$(z_s - z_{s+1})(z'_s - z'_{s+1})\rightarrow$	−.0534	−.0528	−.0424	+.0283	+.0919	+.0285				
$\frac{Nf_{xy}}{f_x f_y}(z_s - z_{s+1})(z'_s - z'_{s+1})\rightarrow$	−.0368	−.0314	−.0319	+.0300	+.1289	+.0511				
	41	100	202	255	138	15	751	.435	.2766	−.1006
	97,630	164,469	305,657	401,785	281,625	44,309				
	.7244	1.0488	1.1400	1.0948	.8453	.5840				
	+.0142	+.0141	+.0113	−.0075	−.0245	−.0076				
	+.0103	+.0148	+.0129	−.0082	−.0207	−.0044				
	39	58	70	61	33	4	265	.154	.0897	−.1869
	34,450	58,035	107,855	141,775	99,375	15,635				
	1.9528	1.7240	1.1196	.7422	.5728	.4413				
	+.0265	+.0262	+.0210	−.0140	−.0455	−.0141				
	+.0517	+.0452	+.0235	−.0104	−.0261	−.0062				
	17	13	22	10	10	1	73	.042	.0000	−.0897
	9,490	15,987	29,711	39,055	27,375	4,307				
	3.0901	1.4027	1.2773	.4417	.6301	.4005				
	+.0127	+.0126	+.0101	−.0067	−.0219	−.0068				
	+.0392	+.0177	+.0129	−.0030	−.0138	−.0027				
f_x	130	219	407	535	375	59	1725			
f_x/N	.0754	.1270	.2359	.3101	.2174	.0342		1.000		
z_s	.0000	.1416	.2816	.3941	.3191	.0755	$\Sigma f_{xy}\frac{N}{f_x f_y}(z_s - z_{s+1})(z'_s - z'_{s+1}) = .2426$			
$z_s - z_{s+1}$	−.1416	−.1400	−.1125	+.0750	+.2436	+.0755				

TABLE 67 (CONTINUED)

$\frac{N}{f_x}$	$(z_s - z_{s+1})^2$	$\frac{N}{f_x}(z_s - z_{s+1})^2$	$\frac{N}{f_y}$	$(z'_s - z'_{s+1})^2$	$\frac{N}{f_y}(z'_s - z'_{s+1})^2$
29.237	.00570	.1667	2.712	.14228	.3859
4.600	.05934	.2730	2.297	.01012	.0232
3.224	.00562	.0181	6.509	.03493	.2274
4.238	.01266	.0537	23.630	.00805	.1902
7.877	.01960	.1544			.8267
13.269	.02005	.2660			
		.9319			

$$c r_{xy} = \frac{.2426}{.9319 \times .8267} = .315$$

By using X for intelligence and Y for quality of clothing, the ordinates on the two scales may be found in the usual way, and the quantities needed for formula (120) may be worked out as shown in Table 67, pp. 264 and 265. Holzinger's Table XII has been used throughout. The corrective value becomes .315. Comparing this result with that obtained in the computation by Professor Pearson, it must be noted that his .317 was worked out with a somewhat different corrective formula.

Needless to say, the arithmetic is very laborious and must be done on a calculator. The above correction, however, is important, and formula (120) or similar forms given in Pearson's paper should be used for the best results.

In case only a rough approximation to the correlation is desired, class values such as 1, 2, 3, · · · may be assigned to both sets of categories, and the coefficient may be worked out by the method of Chapter IX. The student is urged to work out this value for the above problem in order to compare results.

4. The Correlation Ratio for Qualitative and Unordered Series

When a series has been represented on a normal scale, the calculation of the correlation ratio becomes very simple. The work will be illustrated by the problem of the preceding section.

Since $M_x = 0$, formula (61) for the correlation ratio based on the means of the rows becomes

$$\eta_{xy} = \frac{\sqrt{\frac{\Sigma f_y \bar{x}_y^2}{N}}}{\sigma_x} = \sqrt{\frac{\Sigma f_y \left(\frac{\bar{x}_y}{\sigma_x}\right)^2}{N}}. \quad \left\{\begin{array}{c}\text{Correlation ratio adapted}\\ \text{for use with data on a nor-}\\ \text{mal scale}\end{array}\right\} \quad (121)$$

For the data given in Table 64 the arithmetic may be arranged as shown in Table 68. The work is very easily done in this problem because the means of the rows are already worked out in Table 65. The complete calculation is shorter than that for the correlation coefficient, since σ_y is not required.

TABLE 68. ILLUSTRATING THE CALCULATION OF THE CORRELATION RATIO WITH FORMULA (121)

$\frac{\bar{x}_y}{\sigma_x}$	$\left(\frac{\bar{x}_y}{\sigma_x}\right)^2$	f_y	$f_y\left(\frac{\bar{x}_y}{\sigma_x}\right)^2$	
.795	.6320	4	2.5280	
.615	.3782	37	13.9934	$\frac{31.8786}{245} = .13012$
.121	.0146	68	0.9928	
−.129	.0166	84	1.3944	$\therefore \eta = \sqrt{.13012} = .361$
−.345	.1190	40	4.7600	
−.776	.6022	10	6.0220	
−.444	.1971	1	0.1971	
−1.411	1.9909	1	1.9909	
		245	31.8786	

Applying Blakeman's shorter test for linearity, we find that

$$\sqrt{245}\sqrt{(.361)^2 - (.352)^2} = 1.25 < 4.05.$$

Since 1.25 is less than one third of 4.05 and N is fairly large, the regression in this case may be regarded as sensibly linear.

If one of the associated series is quantitative or qualitative and the other unordered, one of the correlation ratios may always be found. Thus, if Y be quantitative, the ratio η_{yx} has the form

$$\eta_{yx} = \frac{\sqrt{\frac{\Sigma f_x (M_y - \bar{Y}_x)^2}{N}}}{\sigma_y} \quad \left\{\begin{array}{l}\text{Correlation ratio for}\\ \text{means of columns}\end{array}\right\} \qquad (62)$$

and is to be regarded as the ratio of two standard deviations, both depending upon Y only. The arrangement of the X categories is clearly a matter of indifference, since it will not affect the numerator or σ_y in the above expression.

An example of a qualitative and an unordered table is furnished by some data from a study by Mr. Tulchin of the Chicago Institute for Juvenile Research. A large number of children were rated by their teachers as of the "annoying," "sympathetic," or "unsympathetic" type, and also classified in five intelligence categories. Inasmuch as the three "attitude" categories do not necessarily come in any order, they furnish an unordered series. The table of frequencies appears as shown on page 268.

TABLE 69. TULCHIN'S DATA ON INTELLIGENCE AND ATTITUDE

INTELLIGENCE	ATTITUDE			TOTAL
	Annoying	Unsympathetic	Sympathetic	
5 Very Superior . . .	5	—	219	224
4 Superior	24	12	1213	1249
3 Normal	105	103	2451	2659
2 Inferior	131	108	1021	1260
1 Very Inferior	73	82	174	329
Total	338	305	5078	5721

Although the method employed with this problem will be the same as that for the physics test, the results will be worked out for the purpose of further illustration and for comparison of the association measured by a later method. The percentage frequencies of the intelligence distribution are 5.8, 22.0, 46.5, 21.8, and 3.9, beginning with the Very Inferior group. The ordinates between the pieces by the method of the preceding section are therefore .1160, .3354, .3224, and .0844 (Holzinger's Table XII). By formula (86), the means of the five pieces under the marginal distribution become

$$\frac{\bar{y}_1}{\sigma_y} = \frac{0 - .1160}{.058} = -2.000, \quad \frac{\bar{y}_2}{\sigma_y} = \frac{.1160 - .3354}{.220} = -0.997,$$

$$\frac{\bar{y}_3}{\sigma_y} = +.028, \quad \frac{\bar{y}_4}{\sigma_y} = +1.092, \quad \text{and} \quad \frac{\bar{y}_5}{\sigma_y} = +2.164.$$

Multiplying these class values by the corresponding frequencies in the columns, the means of the three columns become

$$\frac{\bar{y}_a}{\sigma_y} = -.7001, \quad \frac{\bar{y}_u}{\sigma_y} = -.8383, \quad \text{and} \quad \frac{\bar{y}_s}{\sigma_y} = +.0987,$$

the subscripts referring to the verbal categories. The remainder of the computation is given in Table 70.

There are a number of corrections which may be applied to the correlation ratio to adjust for too coarse or too fine grouping. The correction for broad categories may be illustrated in the case

TABLE 70. ILLUSTRATING THE CALCULATION OF THE CORRELATION RATIO FOR TULCHIN'S DATA

$\frac{\bar{y}_x}{\sigma_y}$	$\left(\frac{\bar{y}_x}{\sigma_y}\right)^2$	f_x	$f_x\left(\frac{\bar{y}_x}{\sigma_y}\right)^2$
+ .0987	.009742	5078	49.470
− .8383	.702747	305	214.338
− .7001	.490140	338	165.667
		5721	429.475

$$\frac{429.475}{5721} = .07507 \quad \therefore \eta_{yx} = \sqrt{.07507} = .274$$

of the data in Table 64 for which $\eta_{yx} = .385$. With ${}_c\eta_{yx}$ denoting the corrected ratio and r_{xc} the correlation of x with its class value, Professor Pearson* has shown that

$${}_c\eta_{yx} = \frac{\eta_{yx}}{r_{xc}}, \quad \left\{\begin{matrix}\text{Correlation ratio corrected} \\ \text{for broad categories}\end{matrix}\right\} \quad (122)$$

where $r_{xc} = \sqrt{\Sigma \frac{N}{f_x}(z_s - z_{s+1})^2}$. $\left\{\begin{matrix}\text{Correlation of a variable} \\ \text{with its class value}\end{matrix}\right\}$ (123)

The computation will therefore be as follows:

TABLE 71. ILLUSTRATING THE CALCULATION WITH FORMULA (122), FOR THE DATA OF TABLE 64

$z_s - z_{s+1}$	$(z_s - z_{s+1})^2$	$\frac{N}{f_x}$	$\frac{N}{f_x}(z_s - z_{s+1})^2$
$z_0 - z_1 = -.2766$	.076508	5.1042	.390512
$z_1 - z_2 = -.1211$	.014665	3.6567	.053626
$z_2 - z_3 = +.0861$	.007413	3.4507	.025580
$z_3 - z_4 = +.3116$	.097095	4.1525	.403187
			.872905

$$r_{xc} = \sqrt{.872905} = .934 \quad \therefore {}_c\eta_{yx} = \frac{.385}{.934} = .412$$

In case there is a fairly large number of categories and N is not large, a correction for *fineness* of grouping may become

* Karl Pearson, "On the Measurement of the Influence of Broad Categories upon Correlation," *Biometrika*, Vol. IX, p. 116. See, also, Student, "The Correction to be made to the Correlation Ratio for Grouping," ibid. p. 316.

important. This adjustment is especially important in dealing with small coefficients even if N be large, as may be illustrated by an example in a paper* by the writer. The correlation ratio for breathing capacity on reaction time to sight was found to be .1404. Mr. R. A. Fisher† has proved that when we sample from material for which the actual value of η is zero, and t is the number of arrays, then the mean value $\overline{\eta}^2$ from sample to sample will be $\frac{t-1}{N-1}$, where N is the size of the sample. In other words, although the true value is zero, the observed value will not be zero, owing to the grouping and to the sampling deviations which must always enter as positive quantities. In the present example $N = 3373$ and $t = 17$, so that $\overline{\eta}^2$ from this formula is .004745, the probable error of which is $.6745\sqrt{\frac{2\,\overline{\eta}^2(1-\overline{\eta}^2)}{N+1}}$, or .001128. The difference, $\eta^2 - \overline{\eta}^2$, may now be written as $.014967 \pm .001128$, and we may conclude that it is extremely unlikely that the ratio found could have arisen from the fluctuations in uncorrelated material.

For breathing capacity on keenness of hearing we find, likewise, $\eta = .0840 \pm .0115$, $t = 15$, and $\eta^2 - \overline{\eta}^2 = .002904 \pm .001056$. In this case the observed value would appear to be significant by the usual test based on its own probable error; but when η^2 and $\overline{\eta}^2$ are compared, their difference is less than three times the probable error of $\overline{\eta}^2$, and hence the observed correlation of .0840 may be ascribed to the fluctuations in sampling. Breathing capacity and keenness of hearing are therefore uncorrelated.

Corrections for coarseness of grouping may also be made in the case of the correlation coefficient. The reader is referred to Sheppard's corrections given in Chapter XVI and to a paper by Professor Pearson.‡

* Karl J. Holzinger, "On the Relation of Vital Capacity to Certain Psychical Characters," *Biometrika*, Vol. XVI, p. 145.

† R. A. Fisher, "The Goodness of Fit of Regression Formulas," *Journal of the Royal Statistical Society*, Vol. 85, p. 597.

‡ Karl Pearson, "On the Correction Necessary for the Correlation Ratio," *Biometrika*, Vol. XIV, p. 412.

5. Biserial r

If one of the characters in a table such as Table 72 is quantitative and the other consists merely of two qualitative categories, it is possible to find the correlation very simply by a method known as biserial r.

In the derivation of this coefficient it is only necessary to assume that the distribution of the twofold (or dichotomous) character is normal, and that the regression in the table is linear. From Fig. 68, where the usual notation is illustrated, it appears at once that the slope of the regression line is given by

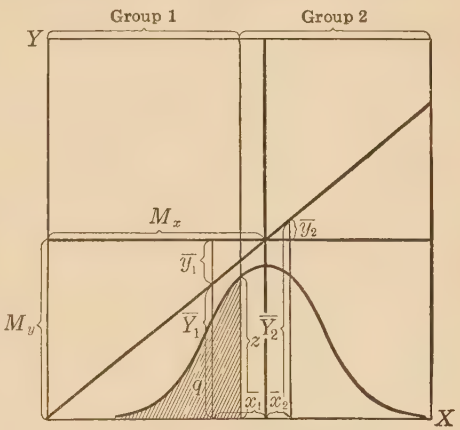


Fig. 68. Illustrating biserial r

$$b_{yx} = \frac{\bar{y}_2}{\bar{x}_2} = \frac{\bar{y}_1}{\bar{x}_1} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1}.$$

Making use of the above value, the correlation coefficient may now be written

$$r = b_{yx}\frac{\sigma_x}{\sigma_y}$$

or

$$r = \frac{\bar{y}_2 - \bar{y}_1}{\sigma_y} \Big/ \frac{\bar{x}_2 - \bar{x}_1}{\sigma_x}.$$

The numerator of this last expression becomes $(\bar{Y}_2 - \bar{Y}_1)/\sigma_y$, that is, the difference between the means of the two columns, divided by the standard deviation of Y for the whole table.

The quantities $\frac{\bar{x}_2}{\sigma_x}$ and $\frac{\bar{x}_1}{\sigma_x}$ are the means of the two pieces under the normal curve and are readily found by the use of formula (86). Denoting the fractional area $\frac{n_1}{N}$ by q and the remaining area $\frac{n_2}{N}$ by p, it follows from (86) that

$$\frac{\bar{x}_2 - \bar{x}_1}{\sigma_x} = \frac{z}{p} + \frac{z}{q} = \frac{z(q+p)}{pq} = \frac{z}{pq}.$$

The desired formula may then be written

$$r_{\text{bis.}} = \frac{\bar{Y}_2 - \bar{Y}_1}{\sigma_y}\left(\frac{pq}{z}\right). \qquad \{\text{Biserial } r\} \qquad (124)$$

TABLE 72. RENT AND HEALTH OF YEARLING BABIES ILLUSTRATING THE METHOD OF BISERIAL r

RENT IN SHILLINGS	HEALTH		TOTAL	
	Not Good (1)	Good (2)		
8.5	1	1	2	$\frac{n_1}{N} = q = .2856$
8.0	—	—	—	
7.5	—	4	4	$\frac{n_2}{N} = p = .7144$
7.0	—	4	4	
6.5	1	13	14	$z = .3399$
6.0	1	18	19	$\bar{Y}_1 = 3.7065$
5.5	4	45	49	$\bar{Y}_2 = 4.1798$
5.0	16	82	98	$\sigma_y = .8021$
4.5	53	252	305	(Sheppard's
4.0	101	303	404	correction)
3.5	132	182	314	
3.0	55	64	119	$r = .354$
2.5	26	18	44	
2.0	7	7	14	
Total	$397 = n_1$	$993 = n_2$	$1390 = N$	

The computation will next be illustrated by Table 72. The means $\bar{Y}_1$ and $\bar{Y}_2$ are found to be 3.7065 and 4.1798, respectively, and $\sigma_y = .8021$ with Sheppard's correction. Next, dividing n_1 by N gives $q = .2856$ and dividing n_2 by N gives $p = .7144$. Upon entering Holzinger's Table XII with $p - .5 = .2144$ the value for z is found to be .3399 with linear interpolation. Substituting all these values in formula (124), we find that

$$r = \frac{(.4733)(.2040)}{(.8021)(.3399)} = \frac{.09655}{.2726} = .3542.$$

We may therefore conclude that there was some tendency for the good health of yearling babies to be associated with a relatively high rent for the home.

The probable error for biserial r when q is not less than .05 is given approximately as

$$P.E._{(\text{bis. } r)} = \frac{.6745\left(\sqrt{\frac{pq}{z^2}} - r^2\right)}{\sqrt{N}}. \quad \left\{\begin{matrix}\text{Probable error} \\ \text{of biserial } r\end{matrix}\right\} \quad (125)$$

6. The Coefficient of Contingency

When both characteristics are unordered the above methods cannot be used, and we must resort to the theory of probability in order to secure a measure of association. To illustrate this method, which is known as *contingency*, we may take a very simple correlation table such as the following, the numbers being taken small for convenience.

	A	B	C	f_y
L		1	2	3
M	2	5		7
N	2	4	1	7
O	3			3
f_x	7	10	3	20

For the cell marked in heavy lines, we shall have

$$f_{xy} = 5, \quad f_x = 10, \quad \text{and} \quad f_y = 7.$$

The probability that a measure will fall in a given column f_x is f_x/N (for example, $\frac{10}{20}$), since f_x of the N equally likely occurrences are favorable. Similarly, the probability that a measure will fall in a particular row is f_y/N (for example, $\frac{7}{20}$). If now these two events are regarded as *independent*, the probability for their combined occurrence is the product of the two probabilities above, or $\frac{f_x f_y}{N^2}$ (for example, $\frac{70}{400}$). Out of the N measures, therefore, we should expect $N\left(\frac{f_x f_y}{N^2}\right)$, or $\frac{f_x f_y}{N}$, to fall in a particular cell if the characters are entirely independent.

For the marked cell the observed frequency is 5 as compared with an independence frequency of $\frac{70}{20}$, or 3.5. The difference $5 - 3.5 = 1.5$, or in general $f_{xy} - \frac{f_x f_y}{N}$, is thus a measure of the departure of the two characters from complete independence, that is, of *contingency.*

Professor Pearson* has defined the mean square contingency for the whole table by the relation

$$\phi^2 = \frac{\chi^2}{N} = \frac{1}{N}\Sigma\left[\frac{\left(f_{xy} - \frac{f_x f_y}{N}\right)^2}{\frac{f_x f_y}{N}}\right] \cdot \left\{\begin{matrix}\text{Mean square}\\ \text{contingency}\\ \text{function}\end{matrix}\right\} \quad (126)$$

The χ^2 function, it will be noted, is the same as that used in Chapter XIII. What is really wanted, however, is a coefficient varying between 0 and 1, and this is given by

$$C = \sqrt{\frac{\phi^2}{1+\phi^2}} = \sqrt{\frac{\chi^2}{N+\chi^2}} \quad \left\{\begin{matrix}\text{Coefficient of}\\ \text{mean square}\\ \text{contingency}\end{matrix}\right\} \quad (127)$$

and called by Pearson the *coefficient of mean square contingency.* In the paper cited in the footnote below he shows that when both of the characters are normally distributed the limiting value for C for many categories is the correlation coefficient r.

A form of (127) which is more convenient for calculation may be obtained by noting that

$$\chi^2 = \Sigma\left\{\frac{f_{xy}^2}{\frac{f_x f_y}{N}}\right\} - 2\,\Sigma f_{xy} + \frac{\Sigma f_x f_y}{N} = S' - N,$$

where S' is the squared sum and N results from the remaining terms. We may therefore write

$$C = \sqrt{\frac{S'-N}{N+S'-N}} = \sqrt{\frac{S'-N}{S'}} \cdot \left\{\begin{matrix}\text{First computa-}\\ \text{tion form for}\\ \text{contingency}\end{matrix}\right\} \quad (128a)$$

* Karl Pearson, "On the Theory of Contingency and its Relation to Association and Normal Correlation," *Draper's Research Memoirs, Biometric Series I,* 1904.

This is the formula recommended by Yule,* but the writer prefers the following one, obtained by setting $S' = NS$.

$$C = \sqrt{\frac{S-1}{S}}, \quad \left\{\begin{matrix}\text{Second computation form for} \\ \text{contingency}\end{matrix}\right\} \quad \textbf{(128b)}$$

where S is now $\Sigma\left\{\frac{f_{xy}^2}{f_x f_y}\right\}$.

The calculation of C is very simple. If formula (128b) is used, the observed cell frequencies f_{xy} are first squared, then the products $f_x f_y$ are obtained, and the quotients $f_{xy}^2/f_x f_y$ worked out. The sum of these last quantities gives S, which may then be substituted in the formula. For the above problem the work may be arranged as follows:

TABLE 73. SHOWING CALCULATION OF THE CONTINGENCY COEFFICIENT C

	A	B	C	f_y
L	f_{xy} f_{xy}^2 $f_x f_y$ $f_{xy}^2/f_x f_y$	1 1 30 .0333	2 4 9 .4444	3
M	2 4 49 .0816	5 25 70 .3571		7
N	2 4 49 .0816	4 16 70 .2286	1 1 21 .0476	7
O	3 9 21 .4286			3
f_x	7	10	3	20

We thus find $S = 1.7028$, whence

$$C = \sqrt{\frac{.7028}{1.7028}} = \sqrt{.4127} = .64.$$

* Yule, Introduction to Statistics, p. 65.

For further illustration and in order to compare the result by this method with that by the correlation ratio, we shall also work out the contingency coefficient for the attitude and intelligence ratings given in Table 69. A table of squares is of course necessary in all such work.

CALCULATION OF THE CONTINGENCY COEFFICIENT FOR TULCHIN'S DATA

	ANNOYING	UNSYMPATHETIC	SYMPATHETIC	f_y
5	5 25 75,712 .000330		219 47,961 1,137,472 .042165	224
4	24 576 422,162 .001364	12 144 380,945 .000378	1213 1,471,369 6,342,422 .231989	1249
3	105 11,025 898,742 .012267	103 10,609 810,995 .013081	2451 6,007,401 13,502,402 .444914	2659
2	131 17,161 425,880 .040295	108 11,664 384,300 .030351	1021 1,042,441 6,398,280 .162925	1260
1	73 5,329 111,202 .047922	82 6,724 100,345 .067009	174 30,276 1,670,662 .018122	329
f_x	338	305	5078	5721

$$S = 1.113112, \quad S - 1 = .113112$$

$$\therefore C = \sqrt{\frac{.113112}{1.113112}} = .319$$

In his text on statistics Mr. Yule* has shown that for t categories each way the contingency coefficient has a maximum value of $\sqrt{\frac{t-1}{t}}$ and that for such a table the largest value for C is given as follows:

* G. Yule, Introduction to Statistics, p. 66.

If $t = 2$, C cannot exceed 0.707.
If $t = 3$, C cannot exceed .816.
If $t = 4$, C cannot exceed .866.
If $t = 5$, C cannot exceed .894.
If $t = 6$, C cannot exceed .913.
If $t = 7$, C cannot exceed .926.
If $t = 8$, C cannot exceed .935.
If $t = 9$, C cannot exceed .943.
If $t = 10$, C cannot exceed .949.

It is well therefore to restrict the use of the coefficient of contingency to 5×5 fold or finer classification whenever possible. For low association values, however, the above difficulty does not enter in any marked degree and the contingency method is always valuable in making a preliminary analysis of a table as illustrated by Professor Pearson in his Tables.* For the example on page 276 the method of contingency is as good as the correlation ratio, and the two results found are in fairly close agreement.

The correction† for broad grouping in the case of the contingency coefficient becomes

$$ {}_cC = \frac{C}{r_{xc}r_{yc}}, \left\{ \begin{array}{c} \text{Correction to the contingency coefficient for broad grouping} \end{array} \right\} \quad (129) $$

where r_{xc} and r_{yc} are given by formula (123). For the problem in Table 66 Professor Pearson finds $C = .291$. The values for r_{xc} and r_{yc} may be easily obtained from the work in Table 67, that is,

$$ r_{xc} = \sqrt{.9319} = .965 \quad \text{and} \quad r_{yc} = \sqrt{.8267} = .909. $$

Substituting these results in formula (129), we find that

$$ {}_cC = \frac{.291}{.965 \times .909} = .332. $$

This is again in close agreement with Pearson's result, ${}_cC = .334$, worked out with another corrective formula (loc. cit. p. 131).

* Pearson's Tables, p. xxxv. † *Biometrika*, Vol. IX, p. 130.

Professor Pearson concludes his paper with the remark that for contingency tables of 5×5 or 6×6 the corrective factors will be small, but for 4×4 or 3×3 tables the corrections are important and should always be made.

The probable error of C is rather awkward to work out. It is given by the formula

$$P.E._C = \frac{.6745}{\sqrt{N}} \left[\frac{\frac{\psi^3}{\phi^2} + 1 - \phi^2}{(1+\phi^2)^3} \right]^{\frac{1}{2}}, \quad \left\{ \begin{matrix} \text{Probable error} \\ \text{of contingency} \\ \text{coefficient} \end{matrix} \right\} \quad (130)$$

where $$\phi^2 = \frac{1}{N} \Sigma \left[\frac{\left(f_{xy} - \frac{f_x f_y}{N} \right)^2}{\frac{f_x f_y}{N}} \right] = S - 1$$

and $$\psi^3 = \frac{1}{N} \Sigma \left[\frac{\left(f_{xy} - \frac{f_x f_y}{N} \right)^3}{\left(\frac{f_x f_y}{N} \right)^2} \right].$$

It is therefore necessary to work out ψ^3 by entering each cell.

7. Correlation from Ranks

When the data are ranked in order of magnitude a rough measure of the correlation is given by Spearman's formula,

$$\rho = 1 - \frac{6 \Sigma (v_x - v_y)^2}{N(N^2 - 1)}, \quad \left\{ \begin{matrix} \text{Spearman's formula} \\ \text{based on rank dif-} \\ \text{ferences} \end{matrix} \right\} \quad (131)$$

where v_x and v_y are the ranks of the X and Y items, respectively.

The above formula may be readily obtained from the product-moment formula by setting $X = v_x$ and $Y = v_y$. By noting that the sum of the squares of the first N integers is given by $N(2N+1)(N+1)/6$, the remainder of the proof may be worked out by forming Σxy, σ_x, and σ_y and is left as an exercise for the student. It may also be shown that ρ ranges in value from -1 to 1.

The calculation of ρ is very simple as shown in the following example, which is limited to 10 cases for illustration. After ranking the items in the two series by the method of Chapter II, the computation may be arranged as shown in Table 74.

TABLE 74. ILLUSTRATING THE CALCULATION OF CORRELATION FROM RANKS

X	Y	v_x	v_y	$(v_x - v_y)$	$(v_x - v_y)^2$
171	117	2	6	− 4	16
169	153	3	1.5	1.5	2.25
128	131	7	4	3	9
141	105	5	7	− 2	4
106	71	9	10	− 1	1
146	130	4	5	− 1	1
87	80	10	9	1	1
114	101	8	8	0	0
187	153	1	1.5	− 0.5	0.25
133	132	6	3	3	9
					43.5

$$\rho = 1 - \frac{6 \times 43.5}{10 \times 99} = 1 - \frac{261}{990} = .74$$

One difficulty in the use of the above formula arises from the fact that a rectilinear form of distribution is assumed, that is, one frequency for each rank. In order to overcome this difficulty Professor Pearson* has given a corrective formula,

$$r = 2 \sin \frac{\pi}{6} \rho, \quad \left\{ \begin{array}{c} \text{Pearson's correction} \\ \text{to Spearman's rank} \\ \text{coefficient} \end{array} \right\} \quad (132)$$

which converts ρ into r under the assumption of a normal distribution. This correction, however, is small, amounting to .018 at most, and is usually not important because lack of normality may introduce an error several times as large as the correction.

The student is urged to make up a short example in which the distributions are very skewed. The correlation coefficient and rank coefficient should be computed and the difference noted.

* Karl Pearson, Mathematical Contributions to Evolution, XVI, p. 12. Cambridge University Press, London.

Another objection to ρ appears when there are a good many ties in rank. This difficulty is illustrated by the following series:

X	Y	v_x	v_y
10	30	5	3
20	30	4	3
30	30	3	3
40	30	2	3
50	30	1	3

If the value for ρ be worked out it becomes .50 instead of zero as found by the product-moment method. The above example is, of course, extreme, but a large proportion of ties in rank will generally be found to produce a correspondingly large error.

When the data are necessarily given in the form of ranks, and when there are not many ties in rank (say less than one fifth of the items), Spearman's rank formula may be conveniently used to give a rough indication of the correlation. While the arithmetic is simple for short series, the ranking and squaring become laborious* beyond 50 cases. The method is, therefore, recommended for about 20 to 40 cases. With more data the product-moment method is theoretically better and more rapid.

The probable error of ρ is given by the formula

$$P.E._{\rho} = \frac{.7063(1-\rho^2)}{\sqrt{N}}, \quad \left\{\begin{array}{c}\text{Probable error of}\\ \text{Spearman's rank}\\ \text{coefficient}\end{array}\right\} \quad (133)$$

from which it appears that ρ is a more unreliable measure of correlation than r.

EXERCISES

1. Work out the product-moment correlation coefficients for the problems of Exercise 1, Chapter IX, using the method illustrated in section 2 of the present chapter. Assuming that the means of the arrays are also needed, compare the total amount of arithmetic with that required by the use of the correlation form.

* A useful table for the calculation of ρ is given in Tables for the Rank Difference Method. The Scott Company Laboratory, Philadelphia, 1920.

2. Work out the contingency coefficient for the following problem:

CORRELATION BETWEEN OCCUPATIONAL STATUS OF PARENT AND NATIVITY OF CHILD

OCCUPATION	NATIVITY					TOTAL
	1	2	3	4	5	
A		8	8	19	15	50
B	4	28	15	51	17	115
C	18	133	26	65	43	285
D	11	73	19	35	12	150
Total	33	242	68	170	87	600

Key: A = Professional Class
B = Merchant
C = Skilled Labor
D = Unskilled Labor

1 = Child born in United States
2 = One parent born in United States
3 = Both parents born in United States
4 = One grandparent born in United States
5 = Both grandparents born in United States

($C = .294$. *Ans.*)

3. Compute the coefficient of contingency for the following table:

CORRELATION BETWEEN NATIVITY AND MENTAL LEVEL OF CHILD

NATIVITY	MENTAL CATEGORY						TOTAL
	F	B	D	N	S	VS	
4				5	10	4	19
3			3	10	12	3	28
2	4	5	24	52	17	5	107
1	2	5	9	31	5		52
Total	6	10	36	98	44	12	206

Key: F = Feeble-minded
B = Border-line
D = Dull
N = Normal
S = Superior
VS = Very Superior

($C = .434$. *Ans.*)

NOTE. The data for Exercises 2 and 3 were furnished by Mrs. Irene Lange.

4. Work out the correlation from ranks for the Otis and Terman test scores of Exercise 1, Chapter II. ($\rho = .7165$. *Ans.*)

Compare the amount of arithmetic with that in the product-moment method.

5. Do the exercise suggested at the bottom of page 278.

6. Compute η_{xy} for the following table.

Health	Nutrition			Total
	C	B	A	
V.R.	4	11	5	20
R.	56	56	12	124
N.	50	49	16	115
R.D.	140	168	37	345
D.	94	89	16	199
V.D.	6	5	1	12
Total . . .	350	378	87	815

($\eta_{xy} = .096$. *Ans.*)

7. Work out C for the data of Exercise 6. ($C = .119$. *Ans.*)

8. Compute r for the data of Exercise 6, using formula (120). ($r = .0768$. *Ans.*)

CHAPTER XV

PARTIAL AND MULTIPLE CORRELATION

1. The Meaning of Partial Correlation

In dealing with correlation thus far the relationship of only the two associated characters has been considered. Each of these, however, is dependent upon many other factors which may influence the observed correlation to a considerable extent. The problem of partial correlation is to find the relationship between two variables when the influence of other variables has been eliminated or when such factors have been held constant.

The conditioning factors may be eliminated by experimental procedure or by the use of a formula as illustrated by the following example. The factors considered are mental age, chronological age, and ossification ratio, the latter being an index of anatomical development based on measurements of the wrist bones. The problem is to discover the relationship between mental and physical development when the influence of age has been eliminated.

Data for the experimental solution of this problem were furnished by records of the Laboratory Schools of The University of Chicago, the work being done by Miss Ethel Abernethy* and others. The children were all measured within a few days of each birthday. In the table given on page 284 it will be noted that not one of these coefficients is significant in comparison with its probable error. We may therefore conclude that for children of the same age, carpal development and mental age are entirely unrelated.

* Ethel M. Abernethy, "Correlation in Physical and Mental Growth," *Journal of Educational Psychology*, October and November, 1925.

TABLE 75. CORRELATION OF MENTAL AGE AND OSSIFICATION RATIO (GIRLS)

CHRONOLOGICAL AGE	NUMBER OF CASES	CORRELATION COEFFICIENT
6–12	120	$+.016 \pm .062$
13	44	$-.137 \pm .100$
14	62	$-.139 \pm .084$
15	29	$-.174 \pm .122$
16	45	$-.022 \pm .101$
17	37	$+.041 \pm .111$

Turning now to the method of partial correlation, we may designate the three variables as follows:

1 = ossification ratio,
2 = mental age,
3 = chronological age.

The correlation between 1 and 2 for 3 fixed is required and is given by the formula

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \cdot \left\{ \begin{matrix} \text{Partial correlation} \\ \text{coefficient for three} \\ \text{variables} \end{matrix} \right\} \quad (134)$$

By taking several hundred cases ranging in age from 5 to 20 years, the three necessary correlations were found to be $r_{12} = .75$, $r_{13} = .87$, and $r_{23} = .83$ (girls). When these values are substituted in the above formula we find that

$$r_{12.3} = \frac{.75 - .87 \times .83}{\sqrt{[1 - (.87)^2][1 - (.83)^2]}} = .101.*$$

For 320 cases the probable error of this result is .037, and for 360 boys we find, similarly, $r_{12.3} = .089 \pm .035$. Neither of these coefficients is three times its probable error, so they are to be regarded as insignificant. The above method then gives results in entire agreement with the experimental procedure of Miss Abernethy. It should be noted that the original correla-

* When a calculating machine is available Miner's Tables for $\sqrt{1 - r^2}$ (Johns Hopkins Press) are most convenient. For logarithmic calculation using Holzinger's Tables, VII, see section 2.

tion of .75 between mental age and ossification ratio is thus due entirely to the correlation of each of these variables with chronological age.

In Chapter IX, section 8, it was shown that some selection lowers the correlation between traits. Thus if a narrow age range were used we should expect altered correlations between ossification ratio and age and between mental and chronological age, with a resulting lower correlation between the physical and mental traits. By restricting the range of age to zero, we reach rigorous selection the effect of which has been noted above. Partial correlation, then, may be regarded as a method for obtaining relationships under rigorous selection of certain conditioning variables.

While it is usually best to isolate factors experimentally it is often not advisable to do so because of the great reduction in the number of cases. The chief factors to be controlled in the above laboratory data are age, sex, and race. If all these are eliminated by selecting the cases, groups of 8 to 15 result, and correlations based on such small numbers are almost worthless. The method of partial correlation makes it possible to use a much larger body of data, eliminating the conditioning factors by means of formulas. It is therefore a very useful and powerful tool in analyzing the relationships in a set of correlated variables.

Partial correlations may be worked out for any number of variables, but the arithmetic beyond four variables becomes very lengthy and tedious. Of the various methods of computation, solution by logarithms is probably best for students who do not have the use of a calculating machine. In the next section we shall therefore give examples of three-variable and four-variable correlations, using logarithms and straight arithmetical substitution.

One important caution to be observed at the outset is to use the method of partial correlation only in case the tables from which the original coefficients are obtained are sensibly linear.

The procedure is then to find the product-moment correlations for all the variables studied, test the tables for linearity by the method of Chapter X, and then substitute the coefficients obtained in suitable formulas, provided the regressions are all sufficiently linear. In case non-linear relationships are found other methods must be employed, such as the procedure described in the last section of Chapter X.

2. Partial Correlation for Three and Four Variables

In dealing with several variables it becomes necessary to use a suitable notation for the various coefficients which arise. If the variables are designated as X_1, X_2, $X_3 \cdots X_n$, the original correlations r_{12}, $r_{13} \cdots r_{23}$, $r_{24} \cdots r_{(n-1)n}$ are known as coefficients of *zero-order*, and the subscripts are called *primary subscripts*.

Correlations such as $r_{12.3}$, $r_{23.1}$, and $r_{12.n}$ are regarded as coefficients of the *first-order*, while the correlations $r_{12.34}$, $r_{23.14}$, and $r_{34.12}$ are said to be coefficients of the *second-order*, and so on. The subscripts following the decimal point are known as *secondary subscripts*.

The general formula for the partial correlation of the order $(n-2)$ for n variables is given by

$$r_{12.34\ldots n} = \frac{r_{12.34\ldots(n-1)} - r_{1n.34\ldots(n-1)}r_{2n.34\ldots(n-1)}}{\sqrt{[1 - r^2_{1n.34\ldots(n-1)}][1 - r^2_{2n.34\ldots(n-1)}]}}. \quad (135)$$

{Partial correlation coefficient of the order $(n-2)$}

This gives the correlation between variables X_1 and X_2 when the remaining $n-2$ variables have been held constant.

Yule* has shown that the order of the secondary subscripts is indifferent, so that $r_{12.34} = r_{12.43}$, and $r_{12.345} = r_{12.354} = r_{12.543}$, etc. These alternative formulas, as we shall see, furnish very useful checks on the arithmetic, since they give independent solutions for the various partial coefficients.

* Yule, Introduction to Statistics, chap. xii. Charles Griffin & Co., London, 1924.

Using formula (135), we may now write down all the possible correlations from three variables. These are evidently

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{[1 - r_{13}^2][1 - r_{23}^2]}}, \tag{136a}$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{[1 - r_{12}^2][1 - r_{23}^2]}}, \tag{136b}$$

and

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{[1 - r_{12}^2][1 - r_{13}^2]}}. \tag{136c}$$

{Partial correlations of first-order}

Similarly, in the case of four variables we shall have

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{[1 - r_{14.3}^2][1 - r_{24.3}^2]}}, \tag{137a}$$

$$r_{12.43} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{[1 - r_{13.4}^2][1 - r_{23.4}^2]}}, \tag{137b}$$

$$r_{13.24} = \frac{r_{13.2} - r_{14.2}r_{34.2}}{\sqrt{[1 - r_{14.2}^2][1 - r_{34.2}^2]}}, \tag{137c}$$

and

$$r_{13.42} = \frac{r_{13.4} - r_{12.4}r_{32.4}}{\sqrt{[1 - r_{12.4}^2][1 - r_{32.4}^2]}}, \tag{137d}$$

etc.

{Partial correlations of second-order}

Since the two primary subscripts may be selected from four in ${}_4C_2 = 6$ ways, there are evidently six possible partial correlations of the second-order with four variables. Each of these six may be obtained in two ways as a check, for example, $r_{12.34} = r_{12.43}$. The total number of arrangements of the subscripts for four variables is therefore twelve. The student should write all these out in full in order to become familiar with the formula and with the notation employed.

As an illustrative problem we shall take some results found by Mr. Cyril Burt. The variables considered may be defined as shown in the list on the following page.

X_1 = mental age on an English revision of the Binet scale,
X_2 = school attainment expressed in educational age,
X_3 = intellectual development as measured in age units by Burt's reasoning test,
X_4 = chronological age.

The observed correlations of zero-order may be arranged as follows:

TABLE 76. BURT'S* INTERCORRELATIONS

	X_1	X_2	X_3
X_1			
X_2	.91		
X_3	.84	.75	
X_4	.83	.87	.70

Burt does not give the tables on which these correlations are based, but we shall assume they are linear and proceed to the calculation of the partial coefficients.

The total number of different correlations of first-order is evidently twelve, since two variables may be selected from four in six ways, each pair furnishing two correlations on account of the interchangeability of the secondary subscripts.

In working out these values it is best to arrange the calculation as in Table 77 so as to identify each step in the computation. In the following work the logarithms of $\sqrt{1-r^2}$ were taken from Holzinger's Table VII and rounded off to four places. Products such as $.84 \times .83 = .6972$ have also been rounded off to three figures (.697), and four-place logarithms used for the remainder of the computation. Greater accuracy than this is unnecessary when the original coefficients are correct to only two places (see Chapter V).

The first item in column (2) is obtained by forming the product $.84 \times .75 = .630$, that is, the product of the coefficients in the first group of three not in line with .91; next, $.910 - .630 = .280$ gives the first entry in column (3); the logarithm of .280 is

* Cyril Burt, Mental and Scholastic Tests, p. 182. King and Son, Ltd., London, 1921.

9.4472 − 10 as shown in column (4). This completes the calculation up to the logarithm of the numerator. The logarithm of the denominator of $r_{12.3}$ is now obtained by adding the logarithms (from Holzinger's Table VII) of $\sqrt{1 - r_{13}^2}$ and $\sqrt{1 - r_{23}^2}$, which are listed in column (1). The first entry in column (5) is then found by adding 9.7345−10 and 9.8205−10, giving 9.5550−10. To complete the calculation for $r_{12.3}$ it is only necessary to subtract the logarithm of the denominator from the logarithm of the numerator (9.8922 − 10 in column (6)), and look up the corresponding number, or anti-logarithm (.780 in column (7)). The remaining correlations are calculated in a similar way.

In finding the coefficients of second-order the first-order values just found may again be arranged in convenient groups of three, and the same scheme of calculation carried out, as illustrated in Table 78. A complete check on the arithmetic is given by two solutions for each second-order coefficient with formulas such as (137). Each of the six second-order values is thus worked out twice, as shown in the table on page 291.

With zero-order coefficients correct to only two places no greater accuracy can be expected in the higher-order coefficients, but three-place values have been used in Table 78, so that the final results may be rounded off to two places.

The interpretation of coefficients such as those found is rendered difficult because of the fact that the first three variables are all measures of the same thing to a certain extent, and holding one or more of them constant gives a result of doubtful meaning. The variables X_1 and X_3 are both measures of intelligence, but $r_{12.34} = +.61$ while $r_{23.14} = -.08$, the latter coefficient being negligible. Burt interprets the coefficient .61 as follows:

"With both age and 'intelligence' (reasoning ability) constant, the partial correlation between school attainments and Binet results remains at .61 · · ·. There can, therefore, be little doubt that with the Binet-Simon scale a child's mental age is a measure not only of the amount of intelligence with which

TABLE 77. SHOWING THE CALCULATION OF CORRELATION COEFFICIENTS OF FIRST-ORDER

(1) Correlation Coefficient of Zero-Order			(2) Product Term of Numerator	(3) Numerator	(4) Logarithm of Numerator	(5) Logarithm of Denominator	(6) Logarithm of Coefficient of First-Order	(7) Correlation Coefficient of First-Order	
Index	Value	Log $\sqrt{1-r^2}$						Index	Value
12	.91	9.6176	.630	.280	9.4472	9.5550	9.8922	12.3	.780
13	.84	9.7345	.683	.157	9.1959	9.4381	9.7578	13.2	.573
23	.75	9.8205	.764	−.014	$\bar{8}$.1461	9.3521	$\bar{8}$.7940	23.1	−.062
12	.91	9.6176	.722	.188	9.2742	9.4394	9.8348	12.4	.684
14	.83	9.7465	.792	.038	8.5798	9.3105	9.2693	14.2	.186
24	.87	9.6929	.755	.115	9.0607	9.3641	9.6966	24.1	.497
13	.84	9.7345	.581	.259	9.4133	9.6003	9.8130	13.4	.650
14	.83	9.7465	.588	.242	9.3838	9.5883	9.7955	14.3	.624
34	.70	9.8538	.697	.003	7.4771	9.4810	7.9961	34.1	.010
23	.75	9.8205	.609	.141	9.1492	9.5467	9.6025	23.4	.400
24	.87	9.6929	.525	.345	9.5378	9.6743	9.8635	24.3	.730
34	.70	9.8538	.653	.047	8.6721	9.5134	9.1587	34.2	.144

TABLE 78. SHOWING THE CALCULATION OF CORRELATION COEFFICIENTS OF SECOND-ORDER

(1) Correlation Coefficient of First-Order			(2) Product Term of Numerator	(3) Numerator	(4) Logarithm of Numerator	(5) Logarithm of Denominator	(6) Logarithm of Coefficient of Second-Order	(7) Correlation Coefficient of Second-Order	
Index	Value	Log $\sqrt{1-r^2}$						Index	Value
12.4	.684	9.8630	.260	.424	9.6274	9.8429	9.7845	12.43	.609
13.4	.650	9.8808	.274	.376	9.5752	9.8251	9.7501	13.42	.562
23.4	.400	9.9621	.445	−.045	$\bar{8}$.6532	9.7438	$\bar{8}$.9094	23.41	−.081
12.3	.780	9.7964	.456	.324	9.5105	9.7276	9.7829	12.34	.607
14.3	.624	9.8929	.569	.055	8.7404	9.6311	9.1093	14.32	.129
24.3	.730	9.8347	.487	.243	9.3856	9.6893	9.6963	24.31	.497
13.2	.573	9.9136	.027	.546	9.7372	9.9879	9.7493	13.24	.561
14.2	.186	9.9924	.083	.103	9.0128	9.9091	9.1037	14.23	.127
34.2	.144	9.9955	.107	.037	8.5682	9.9060	8.6622	34.21	.046
23.1	−.062	9.9992	.005	−.067	$\bar{8}$.8261	9.9384	$\bar{8}$.8877	23.14	−.077
24.1	.497	9.9384	−.001	.498	9.6972	9.9992	9.6980	24.13	.499
34.1	.010	0.0000	−.031	.041	8.6128	9.9376	8.6752	34.12	.047

he is congenitally endowed . . . it is also an index, largely if not mainly, of the mass of scholastic information and skill . . . which he has accumulated in school." (Op. cit. p. 182)

The coefficient $r_{23.14} = -.08$, on the other hand, would seem to show that for children of given chronological and mental ages, reasoning ability (or "intelligence" as Burt calls it) is entirely unrelated to scholastic achievement. Burt and others have claimed that this result shows the reasoning test to be a pure measure of intelligence "independent of schooling." If mental age, however, is a measure of both "intelligence" and achievement, the partial correlation above will necessarily be low because, by fixing X_1, the variables X_2 and X_3 are thereby both restricted. It may also be noted that "schooling" as used here is a measure of relative achievement in school. The fact that the Binet test has higher correlation with such achievement than does Burt's test, indicates that the former is the better guide in predicting scholastic success and is therefore a better intelligence test for practical purposes.

3. Partial Regression Equations for Three Variables

When two variables, X_1 and X_2, are involved it has been shown in Chapter IX that the equation for predicting the most probable value of X_1 for a given value of X_2 is given by the regression equation

$$\bar{X}_1 = r_{12}\frac{\sigma_1}{\sigma_2}X_2 + \text{constant} = b_{12}X_2 + \text{constant}. \qquad (138)$$

{Regression equation for two variables}

This same method of prediction will now be applied to several variables, X_1, X_2, $X_3 \cdots X_n$. The regression equation for estimating X_1 from the remaining $n - 1$ variables is

$$\bar{X}_1 = b_{12.34\cdots n}X_2 + b_{13.24\cdots n}X_3 + \cdots + b_{1\,n.23\cdots(n-1)}X_n + C, \qquad (139)$$

which is known as a linear function of the X's. The quantities $b_{12.34\cdots n}$, $b_{13.24\cdots n}$, $\ldots b_{1\,n.23\cdots(n-1)}$ and C are constants to be

so chosen that the squared differences between the observed values X_1 and the predicted values $\overline{X}_1$ shall be as small as possible; that is, so that $\Sigma(X_1 - \overline{X}_1)^2 =$ a minimum. These differences, $X_1 - \overline{X}_1$, are also known as *errors of estimate* or *residuals*.

By applying the method of least squares in a manner similar to that in Chapter IX, it may be shown that

$$b_{12.34\cdots n} = r_{12.34\cdots n}\frac{\sigma_{1.34\cdots n}}{\sigma_{2.34\cdots n}}, \left\{\begin{array}{l}\text{Regression coefficient}\\ \text{of the order } (n-2)\end{array}\right\} \quad (140)$$

where

$$\sigma_{1.23\cdots n} = \sigma_1\sqrt{(1-r_{12}^2)(1-r_{13.2}^2)\cdots(1-r_{1n.23\cdots(n-1)}^2)}. \quad (141)$$

{Standard deviation of the order $(n-1)$}

The probable error of estimate is of course given by

$$P.E._{est} = .6745\,\sigma_{1.23\cdots n}. \quad (142)$$

{Probable error of estimate}

The value furnished by (140) is known as a partial regression coefficient. It gives the average change in the dependent variable (left member of the equation) for a unit change in the variable to which it is attached, when all the remaining variables are kept constant.

It will be noted that the subscripts are so arranged that the position of any regression coefficient is uniquely determined. Thus for $b_{13.24\cdots n}$ the primary subscript 1 indicates that X_1 is the dependent variable, while 3 shows the variable X_3 to which the coefficient is attached. The remaining secondary subscripts following the decimal point merely show the number of variables involved, and their arrangement is a matter of indifference.

In order to illustrate the above formulas we shall next write out in full the equations for three variables. From equations (140) and (141) we find that

$$\begin{aligned}\overline{X}_1 &= b_{12.3}X_2 + b_{13.2}X_3 + C_1\\ &= r_{12.3}\frac{\sigma_{1.3}}{\sigma_{2.3}}X_2 + r_{13.2}\frac{\sigma_{1.2}}{\sigma_{3.2}}X_3 + C_1\\ &= r_{12.3}\frac{\sigma_1\sqrt{1-r_{13}^2}}{\sigma_2\sqrt{1-r_{23}^2}}X_2 + r_{13.2}\frac{\sigma_1\sqrt{1-r_{12}^2}}{\sigma_3\sqrt{1-r_{32}^2}}X_3 + C_1.\end{aligned}$$

Upon substituting the values or $r_{12.3}$ and $r_{13.2}$, we obtain

$$\bar{X}_1 = \frac{\sigma_1}{\sigma_2}\frac{(r_{12} - r_{13}r_{23})}{1 - r_{23}^2} X_2 + \frac{\sigma_1}{\sigma_3}\frac{(r_{13} - r_{12}r_{23})}{1 - r_{23}^2} X_3 + C_1 \quad (143)$$

and, similarly,

$$\bar{X}_2 = \frac{\sigma_2}{\sigma_1}\frac{(r_{12} - r_{13}r_{23})}{1 - r_{13}^2} X_1 + \frac{\sigma_2}{\sigma_3}\frac{(r_{23} - r_{12}r_{13})}{1 - r_{13}^2} X_3 + C_2 \quad (144)$$

and

$$\bar{X}_3 = \frac{\sigma_3}{\sigma_1}\frac{(r_{13} - r_{12}r_{23})}{1 - r_{12}^2} X_1 + \frac{\sigma_3}{\sigma_2}\frac{(r_{23} - r_{12}r_{13})}{1 - r_{12}^2} X_2 + C_3. \quad (145)$$

{Regression equations for three variables, short form}

The general expression for C is given by

$$C = M_1 - b_{12.34\cdots n}M_2 - b_{13.24\cdots n}M_3 - \cdots - b_{1\,n.23\cdots(n-1)}M_n. \quad (146)$$

{Constant term in regression equation}

When an estimate has been made with a regression equation it is necessary to know something about the reliability of the obtained prediction. The standard error of estimate for predicting X_1 from $n - 1$ other variables is the standard deviation of the residuals given by equation (141) and is interpreted in the same way as $\sigma\sqrt{1 - r^2}$ of Chapter IX. For the above equations we thus find

$$\sigma_{1.23} = \sigma_1\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13.2}^2} = \sigma_1\sqrt{1 - r_{13}^2}\sqrt{1 - r_{12.3}^2},$$

or

$$\sigma_{1.23} = \sigma_1\sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\,r_{12}r_{13}r_{23}}{1 - r_{23}^2}} = \frac{\sigma_1\sqrt{S_{123}}}{\sqrt{1 - r_{23}^2}}, \quad (147)$$

$$\sigma_{2.13} = \sigma_2\sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\,r_{12}r_{13}r_{23}}{1 - r_{13}^2}} = \frac{\sigma_2\sqrt{S_{123}}}{\sqrt{1 - r_{13}^2}}, \quad (148)$$

$$\sigma_{3.12} = \sigma_3\sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\,r_{12}r_{13}r_{23}}{1 - r_{12}^2}} = \frac{\sigma_3\sqrt{S_{123}}}{\sqrt{1 - r_{12}^2}}, \quad (149)$$

{Standard deviations of the second-order}

where S_{123} is the sum of the terms in the numerator of (147). The probable errors of estimate are of course obtained by multiplying the above values by .6745.

In dealing with a three-variable problem for which only zero-order coefficients are required, formulas (143), (144), and (145) will be found convenient. When the partial correlations are available, however, formulas like (139), (140), and (141) will be found much simpler to employ. The computation for the former equations will be illustrated by an example in which success in first-year college work is predicted by the average of four years' work in high school and an intelligence test. The three variables and zero-order coefficients obtained from a sample of 75 cases may be given as follows:

X_1 = criterion of success = average mark for first-year college work.

X_2 = predictor = average mark from four years in high school.

X_3 = predictor = score on the Brown Intelligence Test.

$$M_1 = 78.0\%, \qquad \sigma_1 = 10.21\%, \qquad r_{12} = .666;$$
$$M_2 = 87.2\%, \qquad \sigma_2 = 6.02\%, \qquad r_{13} = .750;$$
$$M_3 = 32.8 \text{ pts.}, \qquad \sigma_3 = 10.35 \text{ pts.}, \qquad r_{23} = .628.$$

It should be noted that the number of cases ($N = 75$) is too small to give very reliable results, but the above example will be used to illustrate the calculation.

By Blakeman's test the correlations all proved to be linear, so that the method of partial correlation is justified in this problem.

The equation required is (143), or

$$\bar{X}_1 = \frac{\sigma_1}{\sigma_2}\frac{(r_{12} - r_{13}r_{23})}{1 - r_{23}^2} X_2 + \frac{\sigma_1}{\sigma_3}\frac{(r_{13} - r_{12}r_{23})}{1 - r_{23}^2} X_3 + C_1,$$

the computation for which may be arranged as in the table on page 296.

Inasmuch as the zero-order values are given to three and four significant figures, a five-place logarithm table has been used. The logarithms of $1 - r^2$ are given directly by Holzinger's Table VI. It will be found necessary to observe the arrangement of the quantities in the formula very carefully in order to combine the proper logarithms.

TABLE 79. SHOWING CALCULATION OF FIRST-ORDER REGRESSION COEFFICIENTS

(1) r		(2) PRODUCT rr	(3) DIFFERENCE $r - rr$	(4) LOG $(r - rr)$	(5) σ		(6) LOG σ	(7) LOG $(1 - r^2)$
12	.666	.4710	.1950	9.29003	1	10.21	1.00903	—
13	.750	.4182	.3318	9.52088	2	6.02	0.77960	—
23	.628	—	—	—	3	10.35	1.01494	9.78220

(8) LOG NUMERATOR [COLS. (4) AND (6)]	(9) LOG DENOMINATOR [COLS. (6) AND (7)]	(10) LOG COEFFICIENT [COLS. (8) AND (6)]	(11) COEFFICIENT OF REGRESSION FIRST-ORDER	
0.29906	0.56180	9.73726	$b_{12.3}$	.5461
0.52991	0.79714	9.73277	$b_{13.2}$	.5405

Using a calculator and Miner's Table for $1 - r^2$, the above computation becomes very much easier:

$$b_{12.3} = \frac{10.21 \times .195}{6.02 \times .605616} = \frac{1.99095}{3.6458} = .5461$$

and

$$b_{13.2} = \frac{10.21 \times .3318}{10.35 \times .605616} = \frac{3.38768}{6.2681} = .5405.$$

The value of C as given by (146) becomes

$$C = 78 - .5461 \times 87.2 - .5405 \times 32.8 = 12.65 \therefore 12.65\%,$$

the unit being the same as for X_1.

Using formula (147), the probable error of estimate may also be worked out from the zero-order coefficients:

$$S_{123} = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\, r_{12} r_{13} r_{23} = .226932.$$

$$\begin{aligned} \log \sqrt{S_{123}} &= 9.67795 \\ \log \sigma_1 &= 1.00903 \\ \log .6745 &= 9.82898 \\ \hline \log \text{prod.} &= 0.51596 \\ \log \sqrt{1 - r_{23}^2} &= 9.89110 \\ \hline \log .6745\, \sigma_{1.23} &= 0.62486 \\ \therefore P.E._{.1.23} &= 4.22\% \end{aligned}$$

The complete regression equation therefore becomes

$$\bar{X}_1 = .546\ X_2 + .540\ X_3 + 12.65\% \pm 4.22\%.$$

It should be noted that the coefficients .546 and .540 may not be compared directly, but that each gives the average change in X_1 for a unit change in X_2 and X_3, respectively, when the other variable is held constant. Thus an increase of 1 per cent in the high-school record is accompanied on the average by an increase of .546 of 1 per cent in the college record, while an increase of one point on the Brown test is accompanied by an increase of .540 of 1 per cent in college standing.

In making a prediction with the above equation it is only necessary to substitute values for X_2 and X_3. A student, for example, may enter the University with a high-school average of 80 and a Brown test score of 40. Upon substituting these values in this last equation the most probable standing of the student in college at the end of the freshman year will be given by 77.93 ± 4.22. It is therefore an even chance that his college rating will be anywhere from 73.71 to 82.15, and the importance of the probable error of estimate is seen in placing a reservation upon the accuracy of the prediction. For a second student with a high-school average of 90 and a Brown score of 50 we find, similarly, $\bar{X}_1 = 88.79 \pm 4.22$. Here it is an even chance that this student's college average will be between 84.57 and 93.01.

The question sometimes arises, Why predict the college standing of students when it is already known in this problem? The standing of only the sample observed is known, however. This criterion is used as a basis for determining the regression equation by means of which predictions may be made with similar groups for which the college standing is unknown. It is assumed, therefore, that other groups will possess the same characteristics as the sample studied so that the equation may also be applied to them. Needless to say, this assumption is often not fulfilled, but the forecast by means of the regression equation is one of the best that can be made on the basis of past experience.

4. Some Cautions in the Use of Regression Equations

Estimates by means of regression equations may fail to be reliable for several reasons:

a. The trend of the data in the observed sample may be imperfectly represented by a linear function. By testing all of the zero-order regressions for linearity this objection may be overcome.

b. Data to which the regression equation is applied may not be comparable with those of the sample from which the equation was derived. This difficulty may be illustrated by the case of a high school with unusually low or high standards of marking. The use of the equation in the last problem in such a case would give misleading results, since the data were obtained from a normal group. It would probably be necessary to work out a separate equation for such schools.

c. The correlations of various orders may be so small that the probable error of estimate becomes relatively large. If this condition prevails predictors having higher correlation with the criterion must be sought, or their number must be increased, as is evident from inspection of formula (141).

d. The number of cases in the sample furnishing the predicting equation may be so small that the regression coefficients are unstable. The probable error of $b_{12.3} = .546$ in the last problem is given by formula (97) of Chapter XIII, that is,

$$P.E._{b_{12.3}} = .6745 \frac{\sigma_{1.23}}{\sigma_{2.3}\sqrt{N}}$$

$$= .6745 \frac{\sigma_1 \sqrt{S_{123}}}{\sigma_2 (1 - r_{23}^2)\sqrt{N}} = .104,$$

or $$b_{12.3} = .546 \pm .104.$$

The value of the regression coefficients based on a very large number of cases might therefore differ considerably from the values actually found in a small sample of 75, which was used here chiefly for numerical illustration.

The above difficulties may be illustrated by an example taken from a study by Dr. W. R. Burgess.* The predicted variable was years of teacher training beyond high school, regression equations for which were formed with time as the predicting factor. In Fig. 69 the data and regression line for one state are shown. By means of the latter, Dr. Burgess predicted that in 1950 the average teacher in Montana would have 1.36 years of training beyond high school.

It should be observed, however, that the data do not furnish a linear trend for the period studied, and the regression line is therefore a bad fit. Furthermore, the forecast has been obtained from a ten-year period and has been projected forty years beyond the range of observation. The assumption that the educational conditions in Montana from 1920 to 1950 will be comparable with those from 1910 to 1920 is unwarranted and the prediction is therefore probably worthless.

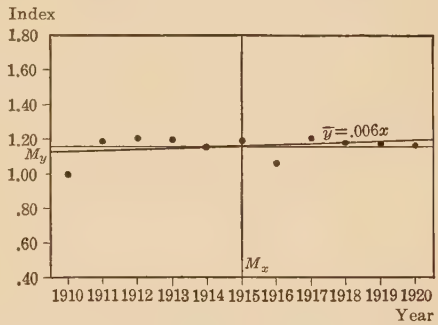


FIG. 69. Regression line for Burgess data

The correlation between teacher-preparation index and time is necessarily small, thus giving a relatively large error in estimate, and finally, although the total number of observations is large, the probable error of a regression coefficient as small as .006 is such as to render its value of doubtful significance.

If predictions of the above type are to be made, the trend for the data studied must be approximately linear and the projection made only a short time beyond the range of observation.

* W. R. Burgess, "Trends of Teacher Preparation," *Journal of Educational Research*, October, 1921, p. 181.

5. Partial Regression Equations for Four Variables

When four variables are involved, the regression equation for predicting X_1 from X_2, X_3, and X_4 may be obtained from formula (139) and written in the form

$$\bar{X}_1 = r_{12.34}\frac{\sigma_{1.34}}{\sigma_{2.34}}X_2 + r_{13.24}\frac{\sigma_{1.24}}{\sigma_{3.24}}X_3 + r_{14.23}\frac{\sigma_{1.23}}{\sigma_{4.23}}X_4 + C.$$

The standard deviations are obtained from (141), giving

$$\begin{aligned}\bar{X}_1 = {} & r_{12.34}\frac{\sigma_1}{\sigma_2}\frac{\sqrt{1-r_{13}^2}\sqrt{1-r_{14.3}^2}}{\sqrt{1-r_{23}^2}\sqrt{1-r_{24.3}^2}}X_2 \\ & + r_{13.24}\frac{\sigma_1}{\sigma_3}\frac{\sqrt{1-r_{12}^2}\sqrt{1-r_{14.2}^2}}{\sqrt{1-r_{23}^2}\sqrt{1-r_{34.2}^2}}X_3 \\ & + r_{14.23}\frac{\sigma_1}{\sigma_4}\frac{\sqrt{1-r_{12}^2}\sqrt{1-r_{13.2}^2}}{\sqrt{1-r_{24}^2}\sqrt{1-r_{34.2}^2}}X_4 + C. \qquad (150)\end{aligned}$$

{Regression equation in four variables in terms of partial correlations}

In order to calculate the regression coefficients for equation (150) it is first necessary to compute the required partial-correlation coefficients of first-order and second-order and then substitute in the above expression for the regression coefficients. The value of the constant term C is then readily determined from formula (146).

This procedure may be the easiest and most direct, especially when the partial correlations are needed for other purposes. Another method will next be presented, however, because students often find it very convenient. The new formulas have two advantages over (150): all the operations involved are fully expressed, and nothing but zero-order correlation coefficients and standard deviations are required in the calculation. It is therefore only necessary to make straightforward substitutions of these values. Since the formulas for the correlation and regression coefficients involve the same expressions, we may begin with the former.

Returning first to the general formula (135) for partial correlation, we may write

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1-r_{14.3}^2)(1-r_{24.3}^2)}}.$$

Upon substituting the values for the coefficients of first-order in this expression, we find

$$r_{12.34} = \frac{r_{12}(1-r_{34}^2) - r_{13}r_{23} - r_{14}r_{24} + r_{34}(r_{13}r_{24} + r_{14}r_{23})}{\sqrt{(1-r_{13}^2-r_{14}^2-r_{34}^2+2r_{13}r_{14}r_{34})(1-r_{23}^2-r_{24}^2-r_{34}^2+2r_{23}r_{24}r_{34})}}$$

$$= \frac{S_{12.34}}{\sqrt{S_{134}S_{234}}}, \tag{151a}$$

and, similarly,

$$r_{13.24} = \frac{r_{13}(1-r_{24}^2) - r_{12}r_{23} - r_{14}r_{34} + r_{24}(r_{12}r_{34} + r_{14}r_{23})}{\sqrt{(1-r_{12}^2-r_{14}^2-r_{24}^2+2r_{12}r_{14}r_{24})(1-r_{23}^2-r_{24}^2-r_{34}^2+2r_{23}r_{24}r_{34})}}$$

$$= \frac{S_{13.24}}{\sqrt{S_{124}S_{234}}}, \tag{151b}$$

and

$$r_{14.23} = \frac{r_{14}(1-r_{23}^2) - r_{12}r_{24} - r_{13}r_{34} + r_{23}(r_{12}r_{34} + r_{13}r_{24})}{\sqrt{(1-r_{12}^2-r_{13}^2-r_{23}^2+2r_{12}r_{13}r_{23})(1-r_{23}^2-r_{24}^2-r_{34}^2+2r_{23}r_{24}r_{34})}}$$

$$= \frac{S_{14.23}}{\sqrt{S_{123}S_{234}}}, \tag{151c}$$

{Second-order correlation coefficients in terms of zero-order coefficients}

where $S_{12.34} = r_{12}(1-r_{34}^2) - r_{13}r_{23} - r_{14}r_{24} + r_{34}(r_{13}r_{24} + r_{14}r_{23})$ and $S_{134} = 1 - r_{13}^2 - r_{14}^2 - r_{34}^2 + 2r_{13}r_{14}r_{34}$, etc., as used in formula (147). It will be noted that seven different expressions of the form indicated by $S_{12.34}$ and S_{134} are required for the computation of the three correlation coefficients.

Similar expressions for the partial regression coefficients are next obtained by the use of a general reduction formula,

$$b_{12.34\cdots n} = \frac{r_{12.34\cdots(n-1)} - r_{1n.34\cdots(n-1)}r_{2n.34\cdots(n-1)}}{1 - r_{2n.34\cdots(n-1)}^2} \frac{\sigma_{1.34\cdots(n-1)}}{\sigma_{2.34\cdots(n-1)}}. \tag{152}$$

{Reduction formula for regression coefficient}

Applying this formula and making use of (141), we find

$$b_{12.34} = \frac{\sigma_1}{\sigma_2}\left[\frac{r_{12.3} - r_{14.3}r_{24.3}}{1 - r_{24.3}^2}\right]\frac{\sqrt{1-r_{13}^2}}{\sqrt{1-r_{23}^2}},$$

and upon substituting the values for first-order correlations, there results

$$b_{12.34} = \frac{\sigma_1}{\sigma_2}\left[\frac{r_{12}(1-r_{34}^2) - r_{13}r_{23} - r_{14}r_{24} + r_{34}(r_{13}r_{24} + r_{14}r_{23})}{1 - r_{23}^2 - r_{24}^2 - r_{34}^2 + 2\,r_{23}r_{24}r_{34}}\right]$$

$$= \frac{\sigma_1}{\sigma_2}\frac{S_{12.34}}{S_{234}}, \qquad (153a)$$

and, similarly,

$$b_{13.24} = \frac{\sigma_1}{\sigma_3}\left[\frac{r_{13}(1-r_{24}^2) - r_{12}r_{23} - r_{14}r_{34} + r_{24}(r_{12}r_{34} + r_{14}r_{23})}{1 - r_{23}^2 - r_{24}^2 - r_{34}^2 + 2\,r_{23}r_{24}r_{34}}\right]$$

$$= \frac{\sigma_1}{\sigma_3}\frac{S_{13.24}}{S_{234}}, \qquad (153b)$$

and

$$b_{14.23} = \frac{\sigma_1}{\sigma_4}\left[\frac{r_{14}(1-r_{23}^2) - r_{12}r_{24} - r_{13}r_{34} + r_{23}(r_{12}r_{34} + r_{13}r_{24})}{1 - r_{23}^2 - r_{24}^2 - r_{34}^2 + 2\,r_{23}r_{24}r_{34}}\right]$$

$$= \frac{\sigma_1}{\sigma_4}\frac{S_{14.23}}{S_{234}}. \qquad (153c)$$

{Second-order regression coefficients in terms of zero-order coefficients}

The advantage of these last equations becomes apparent from the fact that only four different quantities, $S_{12.34}$ and S_{234}, are required for the complete solution of a given regression equation.

The constant term C is of course given by

$$C = M_1 - b_{12.34}M_2 - b_{13.24}M_3 - b_{14.23}M_4.$$

The standard error of estimate may be written

$$\sigma_{1.234} = \sigma_1\sqrt{(1-r_{12}^2)(1-r_{13.2}^2)(1-r_{14.23}^2)}$$

by the use of equation (141). Upon substituting the value for $r_{14.23}$ from (151c) and expressing $r_{13.2}$ in terms of zero-order coefficients, there results, after simplification,

$$\sigma_{1.234} = \sigma_1\sqrt{\frac{S_{123}S_{234} - S_{14.23}^2}{(1-r_{23}^2)S_{234}}}; \quad \left\{\begin{matrix}\text{Standard deviation of}\\ \text{third-order in terms of}\\ \text{zero-order coefficients}\end{matrix}\right\} \qquad (154)$$

and by permuting the subscripts,

$$\sigma_{2.134} = \sigma_2\sqrt{\frac{S_{123}S_{134} - S_{24.13}^2}{(1-r_{13}^2)S_{134}}}, \text{ etc.}$$

The complete solution of the regression equation in four variables, together with the partial correlation coefficients, is thus accomplished by calculating seven quantities, $S_{12.34}$ and S_{134}, based upon zero-order values.

As an illustrative example we may take a four-variable problem worked out by Mr. J. W. Hoge in a term paper. The problem was to predict success in plane geometry from algebraic ability, arithmetical ability, and intelligence. The group consisted of fifty high-school sophomores.

A list of the variables used may be given as follows:

X_1 = criterion = cumulative score on eight units of work in plane geometry covering a six months' period.

X_2 = the cumulative score on three algebra tests covering the four fundamental operations and the solution of linear and quadratic equations.

X_3 = score on the Reavis-Breslich arithmetic test.

X_4 = intelligence quotient on the Otis Self-Administering test.

The zero-order correlation coefficients, standard deviations, and means are given in the following table:

TABLE 80. DATA FROM MR. HOGE'S PAPER

ZERO-ORDER CORRELATIONS	STANDARD DEVIATIONS	MEANS	PARTIAL CORRELATIONS FOR SUBSEQUENT USE
$r_{12} = .54$	$\sigma_1 = 35.5$	$M_1 = 224.4$	
$r_{13} = .49$	$\sigma_2 = 6.87$	$M_2 = 41.32$	$r_{13.2} = .258$
$r_{14} = .41$	$\sigma_3 = 21.28$	$M_3 = 81.52$	$r_{14.23} = .234$
$r_{23} = .58$	$\sigma_4 = 8.49$	$M_4 = 113.88$	
$r_{24} = .29$			
$r_{34} = .50$			

Returning to equations (153), we shall first work out the regression coefficients for the equation

$$\bar{X}_1 = b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 + C.$$

Upon substituting the zero-order correlation coefficients we find $S_{12.34} = .192$, $S_{13.24} = .0779$, $S_{14.23} = .1095$, and $S_{234} = .498$.

The values for the regression coefficients may then be written

$$b_{12.34} = \frac{35.5}{6.87} \times \frac{.192}{.498} = 1.99, \quad b_{13.24} = \frac{35.5}{21.28} \times \frac{.0779}{.498} = .261,$$

and

$$b_{14.23} = \frac{35.5}{8.49} \times \frac{.1095}{.498} = .919.$$

The equation for the constant term C then gives

$$C = 224.4 - 1.99 \times 41.32 - .261 \times 81.52 - .919 \times 113.88 = 16.2,$$

and the complete regression equation is thus written,

$$\bar{X}_1 = 1.99\ X_2 + .261\ X_3 + .919\ X_4 + 16.2.$$

In order to obtain the standard error of estimate $\sigma_{1.234}$, the quantity S_{123} is required, and the latter will be worked out in computing the partial correlation as follows:

$$S_{123} = .439, \quad S_{124} = .585, \quad S_{134} = .543.$$

Substituting the necessary values in equations (151), we find

$$r_{12.34} = \frac{.192}{\sqrt{.543 \times .498}} = .369, \quad r_{13.24} = \frac{.0779}{\sqrt{.585 \times .498}} = .144,$$

and

$$r_{14.23} = \frac{.1095}{\sqrt{.439 \times .498}} = .234.$$

The value for $\sigma_{1.234}$ from formula (154) becomes

$$35.5\sqrt{\frac{.218622 - .011990}{.6636 \times .498}} = 28.1,$$

so that the probable error of estimate is $.6745 \times 28.1 = 19.0$.

This large error of estimate is due to the wide range (134–276) and large standard deviation (35.5) of the cumulative geometry scores as well as to the low intercorrelation of the tests.

By dropping intelligence as a predictor the regression equation becomes $\bar{X}_1 = 1.99\ X_2 + .444\ X_3 + 106.0 \pm 19.5$, with only a slightly larger error of estimate. The estimate from three variables is thus more reliable than the estimate from two variables, but on account of the small difference it is hardly worth while using more than the two predicting variables in

such a problem. Correlations of the order .5 between criterion and predictor are usually necessary before additional variables increase the reliability of the estimate to any appreciable extent.

In order to assist the student in working out any regression coefficients by the above method with four variables, a complete set of values is given in Table 81.

TABLE 81. REGRESSION COEFFICIENTS OF SECOND-ORDER EXPRESSED IN TERMS OF ZERO-ORDER COEFFICIENTS

$$b_{12.34} = \frac{\sigma_1}{\sigma_2}\left[\frac{r_{12}(1-r_{34}^2) - r_{13}r_{23} - r_{14}r_{24} + r_{34}(r_{13}r_{24} + r_{14}r_{23})}{1 - r_{23}^2 - r_{24}^2 - r_{34}^2 + 2\,r_{23}r_{24}r_{34}}\right] = \frac{\sigma_1}{\sigma_2}\frac{S_{12.34}}{S_{234}}$$

$$b_{13.24} = \frac{\sigma_1}{\sigma_3}\left[\frac{r_{13}(1-r_{24}^2) - r_{12}r_{23} - r_{14}r_{34} + r_{24}(r_{12}r_{34} + r_{14}r_{23})}{1 - r_{23}^2 - r_{24}^2 - r_{34}^2 + 2\,r_{23}r_{24}r_{34}}\right] = \frac{\sigma_1}{\sigma_3}\frac{S_{13.24}}{S_{234}}$$

$$b_{14.23} = \frac{\sigma_1}{\sigma_4}\left[\frac{r_{14}(1-r_{23}^2) - r_{12}r_{24} - r_{13}r_{34} + r_{23}(r_{12}r_{34} + r_{13}r_{24})}{1 - r_{23}^2 - r_{24}^2 - r_{34}^2 + 2\,r_{23}r_{24}r_{34}}\right] = \frac{\sigma_1}{\sigma_4}\frac{S_{14.23}}{S_{234}}$$

$$b_{21.34} = \frac{\sigma_2}{\sigma_1}\left[\frac{r_{12}(1-r_{34}^2) - r_{23}r_{13} - r_{24}r_{14} + r_{34}(r_{23}r_{14} + r_{24}r_{13})}{1 - r_{13}^2 - r_{14}^2 - r_{34}^2 + 2\,r_{13}r_{14}r_{34}}\right] = \frac{\sigma_2}{\sigma_1}\frac{S_{12.34}}{S_{134}}$$

$$b_{23.14} = \frac{\sigma_2}{\sigma_3}\left[\frac{r_{23}(1-r_{14}^2) - r_{12}r_{13} - r_{24}r_{34} + r_{14}(r_{12}r_{34} + r_{24}r_{13})}{1 - r_{13}^2 - r_{14}^2 - r_{34}^2 + 2\,r_{13}r_{14}r_{34}}\right] = \frac{\sigma_2}{\sigma_3}\frac{S_{23.14}}{S_{134}}$$

$$b_{24.13} = \frac{\sigma_2}{\sigma_4}\left[\frac{r_{24}(1-r_{13}^2) - r_{12}r_{14} - r_{23}r_{34} + r_{13}(r_{12}r_{34} + r_{23}r_{14})}{1 - r_{13}^2 - r_{14}^2 - r_{34}^2 + 2\,r_{13}r_{14}r_{34}}\right] = \frac{\sigma_2}{\sigma_4}\frac{S_{24.13}}{S_{134}}$$

$$b_{31.24} = \frac{\sigma_3}{\sigma_1}\left[\frac{r_{13}(1-r_{24}^2) - r_{23}r_{12} - r_{34}r_{14} + r_{24}(r_{23}r_{14} + r_{34}r_{12})}{1 - r_{12}^2 - r_{14}^2 - r_{24}^2 + 2\,r_{12}r_{14}r_{24}}\right] = \frac{\sigma_3}{\sigma_1}\frac{S_{13.24}}{S_{124}}$$

$$b_{32.14} = \frac{\sigma_3}{\sigma_2}\left[\frac{r_{23}(1-r_{14}^2) - r_{13}r_{12} - r_{34}r_{24} + r_{14}(r_{13}r_{24} + r_{34}r_{12})}{1 - r_{12}^2 - r_{14}^2 - r_{24}^2 + 2\,r_{12}r_{14}r_{24}}\right] = \frac{\sigma_3}{\sigma_2}\frac{S_{23.14}}{S_{124}}$$

$$b_{34.12} = \frac{\sigma_3}{\sigma_4}\left[\frac{r_{34}(1-r_{12}^2) - r_{13}r_{14} - r_{23}r_{24} + r_{12}(r_{13}r_{24} + r_{23}r_{14})}{1 - r_{12}^2 - r_{14}^2 - r_{24}^2 + 2\,r_{12}r_{14}r_{24}}\right] = \frac{\sigma_3}{\sigma_4}\frac{S_{34.12}}{S_{124}}$$

$$b_{41.23} = \frac{\sigma_4}{\sigma_1}\left[\frac{r_{14}(1-r_{23}^2) - r_{24}r_{12} - r_{34}r_{13} + r_{23}(r_{24}r_{13} + r_{34}r_{12})}{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\,r_{12}r_{13}r_{23}}\right] = \frac{\sigma_4}{\sigma_1}\frac{S_{14.23}}{S_{123}}$$

$$b_{42.13} = \frac{\sigma_4}{\sigma_2}\left[\frac{r_{24}(1-r_{13}^2) - r_{14}r_{12} - r_{34}r_{23} + r_{13}(r_{14}r_{23} + r_{34}r_{12})}{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\,r_{12}r_{13}r_{23}}\right] = \frac{\sigma_4}{\sigma_2}\frac{S_{24.13}}{S_{123}}$$

$$b_{43.12} = \frac{\sigma_4}{\sigma_3}\left[\frac{r_{34}(1-r_{12}^2) - r_{14}r_{13} - r_{24}r_{23} + r_{12}(r_{14}r_{23} + r_{24}r_{13})}{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\,r_{12}r_{13}r_{23}}\right] = \frac{\sigma_4}{\sigma_3}\frac{S_{34.12}}{S_{123}}$$

Another example of four-variable regression is furnished by Burt's data in section 2. The equation for estimating the Binet score from the remaining scores is given by Burt* in the form Binet = .54 school work + .33 intelligence (reasoning) + .11 age, the variables being taken from the mean of the whole set and

* Op. cit. p. 183

all expressed in age units. A year's increase in educational age is therefore accompanied on the average by .54 of a year of increase in mental age, and a year's increase in intelligence by .33 of a year in mental age, etc. Since all the variables are in the same unit and the total of the coefficients happens to be almost unity, Burt makes the following interpretation: "Of the gross result, then, one ninth is attributable to age, one third to intellectual development and over half to school attainment, . . . or in determining the child's performance on the Binet-Simon scale, intelligence can bestow but little more than half the share of school, and age but one third the share of intelligence" (op. cit. p. 183).

These results have been seized upon by the antagonists of intelligence tests as showing that the Binet scale measures chiefly school work and not intelligence, as already noted in section 2; but the difficulties involved in such interpretation become apparent when other equations such as that for predicting age are given. Thus, age = .15 Binet + .51 school work + .03 intelligence.* Are we to conclude from this result that over half a child's age is "attributable" to school work, one sixth to Binet, and only a small fraction to intelligence? Such a conclusion is absurd, but it is logically as sound as Burt's inference regarding his equation.

Regression coefficients of any order merely show the average change in the dependent variable for a unit change in the independent variable to which they are attached, the remaining variables being constant. If these coefficients are obtained for a set of variables all in the same units the relative value of the several predictors may be compared as they affect the estimate. Thus "school work" is five thirds as valuable as "intelligence" in forecasting mental age, and five times as valuable as

* For the derivation of this equation and other critical comment see Holzinger and Freeman, "The Interpretation of Burt's Regression Equation," *Journal of Educational Psychology*, December, 1925. For further discussion see also G. H. Thomson, "The Interpretation of Burt's Regression Equation," and Holzinger and Freeman, "Rejoinder," *Journal of Educational Psychology*, May and September, 1926.

chronological age used for the same purpose. This is quite a different interpretation, however, from Burt's, in which regression coefficients are regarded as representing the parts of the independent variables going to make up the total of the dependent variable.

It must be kept in mind that ordinarily only a few of the possible predicting variables are used in an estimate; also, that the regression coefficients will change whenever a new variable is added, provided the partial correlations involved are not zero. This may be illustrated by beginning with the equation for Binet scores on school work, which is approximately of the form: Binet = .9 school work. According to Burt's reasoning, nine tenths of Binet will then be "attributed to school work." The addition of new variables, however, will reduce this share to .7, then to .54, and much lower if enough predictors are taken which have some partial correlation with Binet.

6. Multiple Correlation

The correlation between a set of observed values such as given by X_1 and the predicted values from the regression equation

$$\bar{X}_1 = b_{12.34\cdots n}X_2 + b_{13.24\cdots n}X_3 + \cdots + b_{1n.23\cdots(n-1)}X_n + C_1$$

is known as multiple correlation and is denoted by $R_{1(234\cdots n)}$. It may be shown* that

$$R_{1(23\cdots n)} = r_{X_1\bar{X}_1} = \sqrt{1 - \frac{\sigma^2_{1.23\cdots n}}{\sigma^2_1}}, \quad \left\{\begin{matrix}\text{Multiple-correlation}\\ \text{coefficient}\end{matrix}\right\} \quad (155)$$

but a more convenient form for calculation is given by

$$1 - R^2_{1(23\cdots n)} = (1 - r^2_{12})(1 - r^2_{13.2})(1 - r^2_{14.23})\cdots(1 - r^2_{1n.23\cdots(n-1)}), \quad (156)$$

$$\{\text{Computation form for } R\}$$

which follows at once from equation (141).

* Yule, op. cit. p. 248.

One use of multiple correlation is in showing how closely X_1 can be expressed as a linear function of X_2, $X_3 \cdots X_n$. If X_1 coincides with the predicted $\overline{X}_1$ for all the observations, the standard error of estimate becomes zero and $R_{1(23\cdots n)}$ by formula (155) will equal unity. If, on the other hand, the residuals $X_1 - \overline{X}_1$ are so large that their standard deviation $\sigma_{1.23\cdots n}$ approaches σ_1, the value of $R_{1(23\cdots n)}$ will approach zero. The multiple-correlation coefficient thus gives an alternative method for determining the reliability of an estimate from a regression equation.

In order to illustrate the use of these formulas, we may return to the data given in section 5 for predicting success in geometry. Considering that X_1 is to be estimated from X_2 and X_3, we may substitute $r_{12} = .54$ and $r_{13.2} = .258$ in equation (156), giving

$$1 - R^2_{1(23)} = [1 - (.54)^2][1 - (.258)^2].$$

The calculation is very easily done with the aid of Holzinger's Table VI, thus:

$$\begin{aligned} \log [1 - \ (.54)^2] &= 9.85028 \\ \underline{\log [1 - (.258)^2] = 9.97008} \\ \log [1 - R^2_{1(23)}] &= 9.82036 \\ \therefore \ R_{1(23)} &= .582. \end{aligned}$$

It is only necessary to add the logarithms and look up the value for R corresponding to their sum; for example, for the logarithm 9.82038 in Table VI we obtain $R = .582$, the answer being correct to three places.

In estimating X_1 from X_2, X_3, and X_4 the equation will be $1 - R^2_{1(234)} = (1 - r^2_{12})(1 - r^2_{13.2})(1 - r^2_{14.23})$. The necessary arithmetic is therefore

$$\begin{aligned} \log [1 - \ (.54)^2] &= 9.85028 \\ \log [1 - (.258)^2] &= 9.97008 \\ \underline{\log [1 - (.234)^2] = 9.97554} \\ \log [1 - R^2_{1(234)}] &= 9.79590 \\ \therefore \ R_{1(234)} &= .612. \end{aligned}$$

When $\sigma_{1.234}$ has already been computed by formula (154), it is of course necessary only to substitute this result in formula (155).

The regression coefficients are the best possible weights which can be assigned to the variables X_2, $X_3 \cdots X_n$ in making a linear prediction for X_1. The multiple-correlation coefficient, therefore, gives a useful measure of the correlation which can be expected from pooling the predictive tests in the form of a regression equation. Thus in the above example the coefficients .582 and .612 measure the reliability of estimates from pooling two and three predictors in the best linear form. The gain in reliability is very slight, however, when a third variable is added, a conclusion which was reached also by comparing the probable errors of estimate, 19.5 and 19.0.

An interesting application of the method of multiple correlation is given in the volume on Psychological Testing in the United States Army,* where the possibility of increasing the correlation between the Beta scale and the Stanford-Binet test is determined. The necessary zero-order coefficients are given in the following table.

TABLE 82. CORRELATIONS OF BETA TESTS WITH STANFORD-BINET MENTAL AGE AND WITH EACH OTHER (653 CASES)

TEST	BETA TESTS							
	1	2	3	4	5	6	7	8
Stanford-Binet	.465	.545	.614	.639	.622	.586	.610	.572
Beta Tests								
1. Maze		.477	.522	.514	.457	.490	.510	.476
2. Cube			.632	.576	.560	.556	.592	.551
3. X–O series				.689	.670	.584	.597	.619
4. Digit symbol					.766	.654	.584	.695
5. Number check						.619	.521	.703
6. Picture							.555	.569
7. Geometrical								.559
8. Spot pattern								

* Memoirs of the National Academy of Sciences, Vol. XV (1921), p. 387.

Upon applying the multiple-correlation formula (155) it was found that $R_{s(12345678)} = .731$, which is the highest correlation obtainable between Stanford-Binet and the best linear weighting of the eight Beta tests. The correlation between the unweighted pool of these eight tests and the Binet test was .728, showing that very slight improvement is made by weighting such components.

The writers of the above report then decided to eliminate certain tests as suggested by the results from the partial correlations and thus obtain a shorter and possibly as good a test with unweighted items as with the whole battery weighted or unweighted. By empirical trial they found: (1) elimination of test 8, $r(\text{Stanford} \times \text{Beta}) = .726$; (2) elimination of tests 8 and 2, $r(\text{Stanford} \times \text{Beta}) = .723$; (3) elimination of tests 8, 2, and 1, $r(\text{Stanford} \times \text{Beta}) = .723$. Thus the simple pool of five of the Beta tests gave almost as good results as the best weighting of all eight.

The final form suggested was to use a non-weighted pool of six of the Beta tests, dropping test 8 and giving test 1 one half the weight of the rest. The correlation for this last result with the Binet scale was .727, which is only slightly less than the best value, .731.

Some important properties of multiple correlation may next be shown by returning to equation (156). It is apparent that every parenthesis on the right is smaller than unity, provided none of the partial correlations be equal to zero. Hence

$$1 - R^2_{1(23\ldots n)} < 1 - r^2_{12},$$
$$1 - R^2_{1(23\ldots n)} < 1 - r^2_{13.2},$$
and $$1 - R^2_{1(23\ldots n)} < 1 - r^2_{14.23}, \text{ etc.}$$
Similarly, $$1 - R^2_{1(32\ldots n)} < 1 - r^2_{13},$$
and $$1 - R^2_{1(42\ldots n)} < 1 - r^2_{14}, \text{ etc.}$$

The multiple-correlation coefficient R cannot, therefore, be smaller than any partial coefficient of zero or of a higher order, and it is usually considerably larger.

If the coefficients $r_{12}, r_{13} \cdots r_{1n}$ are all equal and are denoted by r_{1x}, and if the coefficients $r_{23}, r_{24} \cdots r_{(n-1)n}$ are also equal and are denoted by r_{xx}, it also follows from (155) that

$$R_{1(23\ldots n)} = r_{1x}\sqrt{\frac{n-1}{1+(n-2)r_{xx}}}. \quad \left\{\begin{array}{c}\text{Multiple-correlation}\\ \text{coefficient for equal}\\ \text{coefficients}\end{array}\right\} \quad (157)$$

In case C is to be predicted from n other variables, $n-1$ may be replaced by n in formula (157), giving

$$R_{c(123\ldots n)} = r_{cx}\sqrt{\frac{n}{1+(n-1)r_{xx}}}, \quad (158)$$

which is the same result as that obtained in equation (51) of Chapter IX.

If the numerator and denominator under the radical of equation (158) be divided by n, and then n be allowed to approach infinity, we find that

$$R_{c(123\ldots\infty)} \doteq \frac{r_{cx}}{\sqrt{r_{xx}}}. \quad \left\{\begin{array}{c}\text{Limiting value for}\\ (158)\ \text{when}\ n\to\infty\end{array}\right\} \quad (159)$$

These last two equations are useful in estimating the limits for prediction. Suppose, for example, that there are 50 unrelated environmental conditions, each correlated to the extent of .05 with human physical traits ($r_{xx} = 0$, and $r_{cx} = .05$). Upon substituting in (158), we find $R = .05\sqrt{50} = .35$. In actual practice, however, there is a correlation of about .5 between such environmental conditions, so that by using (159) we find $R \doteq .07$; that is, an infinity of such conditions increase the correlation from .05 to only .07.

The best results are of course obtained by seeking predictors which correlate high with the criterion and low amongst themselves. Thus, if $r_{cx} = .6$, $r_{xx} = .4$, and $n = 10$, we find, from equation (158), that $R = .88$. Arbitrary values may be substituted in this formula, giving a result greater than unity; but, from the constitution of the whole set of variables, this cannot occur in actual practice.

7. Solution by Determinants

While the methods of calculation shown thus far are probably as convenient as any up to four variables, another procedure will next be given in which determinants are employed. The student who is familiar with the theory of determinants and who has the use of a calculating machine may find this method fairly rapid.

The chief function used is the determinant of all the zero-order correlations given by

$$\Delta = \begin{vmatrix} r_{11} & r_{21} & r_{31} & \cdots & r_{n1} \\ r_{12} & r_{22} & r_{32} & \cdots & r_{n2} \\ r_{13} & r_{23} & r_{33} & \cdots & r_{n3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{1n} & r_{2n} & r_{3n} & \cdots & r_{nn} \end{vmatrix} \cdot \left\{ \begin{matrix} \text{Determinant} \\ \text{of zero-order} \\ \text{coefficients} \end{matrix} \right\} \quad (160)$$

A *minor* such as Δ_{12} is obtained by striking out all the coefficients in the row and column common to r_{12}. A *cofactor*, A_{ij}, is equal to the minor Δ_{ij} with the sign that would be attached in expanding the determinant. Thus the three-rowed determinant

$$\Delta = \begin{vmatrix} r_{11} & r_{21} & r_{31} \\ r_{12} & r_{22} & r_{32} \\ r_{13} & r_{23} & r_{33} \end{vmatrix} \quad \left\{ \begin{matrix} \text{Determinant for} \\ \text{three variables} \end{matrix} \right\} \quad (161)$$

may be written $\Delta = r_{11}\Delta_{11} - r_{12}\Delta_{12} + r_{13}\Delta_{13}$

or
$$\begin{aligned} \Delta &= r_{11}A_{11} + r_{12}A_{12} + r_{13}A_{13} \\ &= r_{11}\begin{vmatrix} r_{22} r_{32} \\ r_{23} r_{33} \end{vmatrix} + r_{12}(-1)\begin{vmatrix} r_{21} r_{31} \\ r_{23} r_{33} \end{vmatrix} + r_{13}\begin{vmatrix} r_{21} r_{31} \\ r_{22} r_{32} \end{vmatrix} \\ &= r_{11}(r_{22}r_{33} - r_{23}^2) + r_{12}(r_{13}r_{23} - r_{12}r_{33}) \\ &\qquad + r_{13}(r_{12}r_{23} - r_{13}r_{22}). \end{aligned}$$

Simplifying this last expression, we find that

$$\Delta = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\, r_{12}r_{13}r_{23}, \quad \left\{ \begin{matrix} \text{Expanded} \\ \text{value of (161)} \end{matrix} \right\} \quad (162)$$

which, of course, is the same as S_{123} of section 5.

With a similar notation Professor Pearson* has shown that

$$r_{12.34\ldots n} = \frac{-A_{12}}{\sqrt{\Delta_{11}\Delta_{22}}}. \tag{163}$$

$$r_{1\,k.34\ldots k\ldots n} = \frac{-A_{1k}}{\sqrt{\Delta_{11}\Delta_{kk}}}. \tag{164}$$

$$R_{1(23\ldots n)} = \sqrt{1 - \frac{\Delta}{\Delta_{11}}}. \tag{165}$$

$$\sigma_{1.23\ldots n} = \sigma_1\sqrt{1 - R^2} = \sigma_1\sqrt{\frac{\Delta}{\Delta_{11}}}. \tag{166}$$

$$b_{12.34\ldots n} = \frac{-A_{12}}{\Delta_{11}}\frac{\sigma_1}{\sigma_2}. \tag{167}$$

$$b_{1\,k.23\ldots k\ldots n} = \frac{-A_{1k}}{\Delta_{11}}\frac{\sigma_1}{\sigma_k}. \tag{168}$$

These formulas will next be illustrated by the problem in predicting geometrical success. Arranging the zero-order coefficients from Table 80 in the form of a determinant, we have

$$\Delta = \begin{vmatrix} 1 & .54 & .49 & .41 \\ .54 & 1 & .58 & .29 \\ .49 & .58 & 1 & .50 \\ .41 & .29 & .50 & 1 \end{vmatrix}.$$

This may be worked out by reducing to a determinant of lower order.

Multiplying each row by the reciprocals of the items in the first columns, we have

Reciprocal of Column 1					Column 1
1	$\Delta =$	1	0.54	0.49	0.41	$\times$ 1.000
1.852		1	1.852	1.074	0.537	$\times$ .54
2.041		1	1.184	2.041	1.020	$\times$ .49
2.439		1	0.707	1.220	2.439	$\times$ .41

(If all the elements of a row (or column) are multiplied by the same number n, the determinant is multiplied by n.)

* Unpublished lecture notes.

Next, subtract the elements of the first row from those of each of the other three rows (this leaves the value of Δ unchanged).

$$\Delta = \begin{vmatrix} 1 & 0.54 & 0.49 & 0.41 \\ 0 & 1.312 & 0.584 & 0.127 \\ 0 & 0.644 & 1.551 & 0.610 \\ 0 & 0.167 & 0.730 & 2.029 \end{vmatrix} \times .1085 = \begin{vmatrix} 1.312 & 0.584 & 0.127 \\ 0.644 & 1.551 & 0.610 \\ 0.167 & 0.730 & 2.029 \end{vmatrix} \times .1085$$

$$= .1085[1.312(1.551 \times 2.029 - .73 \times .61) - .644(.584 \times 2.029 - .127 \times .73) + .167(.584 \times .61 - .127 \times 1.551)] = .3112.$$

The determinant can of course be reduced to two rows before expanding, but the arithmetic from the three-rowed value above is very rapid on a machine.

The other determinants required may be worked out in a similar way, that is,

$$\Delta_{11} = \begin{vmatrix} 1 & .58 & .29 \\ .58 & 1 & .50 \\ .29 & .50 & 1 \end{vmatrix} = +.498, \qquad \Delta_{12} = \begin{vmatrix} .54 & .49 & .41 \\ .58 & 1 & .50 \\ .29 & .50 & 1 \end{vmatrix} = +.192.$$

Also, $\Delta_{22} = +.543$, $\Delta_{33} = +.585$, $\Delta_{44} = +.439$, $\Delta_{13} = -.0779$, and $\Delta_{14} = +.1095$, so that $A_{12} = -.192$, $A_{13} = -.0779$, and $A_{14} = -.1095$.

Substituting these values in formulas (163) to (168), we find

$$r_{12.34} = \frac{+.192}{\sqrt{.498 \times .543}} = .369,$$

$$r_{13.24} = \frac{+.0779}{\sqrt{.498 \times .585}} = .144,$$

$$r_{14.23} = \frac{+.1095}{\sqrt{.498 \times .439}} = .234,$$

$$R_{1(234)} = \sqrt{1 - \frac{.3112}{.498}} = .612,$$

and

$$\sigma_{1.234} = 35.5\sqrt{\frac{.3112}{.498}} = 28.1.$$

$$\therefore \; P.E._{1.234} = .6745\,\sigma_{1.234} = 19.0.$$

Also,
$$b_{12.34} = \frac{+.192}{.498}\frac{35.5}{6.87} = 1.99,$$
$$b_{13.24} = \frac{+.0779}{.498}\frac{35.5}{21.28} = .261,$$
and
$$b_{14.23} = \frac{+.1095 \times 35.5}{.498 \times 8.49} = .919.$$

It should be noted that the above results agree with those found in section 5, since $\Delta_{12} = S_{12.34}$, $\Delta_{13} = -S_{13.24}$, $\Delta_{14} = S_{14.23}$, $\Delta_{11} = S_{234}$, $\Delta_{22} = S_{134}$, $\Delta_{33} = S_{124}$, and $\Delta_{44} = S_{123}$.

When more than four variables are involved, it is probably best to use reduction formulas of the type

$$r_{12.345} = \frac{r_{12.34} - r_{15.34}r_{25.34}}{\sqrt{[1 - r^2_{15.34}][1 - r^2_{25.34}]}} \quad \left\{\begin{matrix}\text{Partial cor-}\\ \text{relation of}\\ \text{third-order}\end{matrix}\right\} \quad (169)$$

and
$$b_{12.345} = \frac{(r_{12.34} - r_{15.34}r_{25.34})}{(1 - r^2_{25.34})}\frac{\sigma_1}{\sigma_2}\frac{\sqrt{1 - r^2_{13}}}{\sqrt{1 - r^2_{23}}}\frac{\sqrt{1 - r^2_{14.3}}}{\sqrt{1 - r^2_{24.3}}}, \quad (170)$$

{Regression coefficient of third-order}

and carry out the arithmetic on a calculating machine with the aid of Miner's Tables. The computation is not only easier than by determinants, but the checks $r_{12.34} = r_{12.43}$ etc. already noted can be conveniently made. A good example of a correlation problem in five variables is given in Pearl's "Medical Statistics and Biometry," p. 329, while other methods of calculation may be found in Kelley's "Statistical Method," chap. xi.

EXERCISES

1. Data: 113 pupils (67 boys and 46 girls). Variables, (1) age, (2) weight, (3) standing height, (4) sitting height. Correlations, $r_{12} = .75$, $r_{13} = .85$, $r_{14} = .79$, $r_{23} = .89$, $r_{24} = .90$, $r_{34} = .94$.

Work out the partial correlations of the second-order.

($r_{12.34} = -.007$, $r_{13.24} = .50$, $r_{14.23} = -.04$,
$r_{23.14} = .26$, $r_{24.13} = .41$, $r_{34.12} = .63$.
Ans.)

2. Calculate the first-order and second-order partial-correlation coefficients from the following data:

$r_{12} = .78, \quad r_{13} = .45, \quad r_{14} = .40, \quad r_{23} = .48, \quad r_{24} = .29, \quad r_{34} = .52.$

$$\left.\begin{array}{l} r_{12.3} = .720 \\ r_{12.4} = .757 \end{array}\right\} r_{12.34} = .73 \qquad \left.\begin{array}{l} r_{23.1} = .231 \\ r_{23.4} = .403 \end{array}\right\} r_{23.14} = .27$$

$$\left.\begin{array}{l} r_{13.2} = .138 \\ r_{13.4} = .309 \end{array}\right\} r_{13.24} = .01 \qquad \left.\begin{array}{l} r_{24.1} = -.038 \\ r_{24.3} = .054 \end{array}\right\} r_{24.13} = -.15 \qquad \left.\begin{array}{l} \\ \\ \end{array}\right\} Ans.$$

$$\left.\begin{array}{l} r_{14.2} = .291 \\ r_{14.3} = .218 \end{array}\right\} r_{14.23} = .26 \qquad \left.\begin{array}{l} r_{34.1} = .415 \\ r_{34.2} = .454 \end{array}\right\} r_{34.12} = .44$$

3. Given: $r_{12} = -.481, \quad r_{13} = -.697, \quad r_{14} = -.494,$
$r_{23} = +.374, \quad r_{24} = +.363, \quad r_{34} = +.286,$

$\sigma_1 = 34.48, \quad \sigma_2 = 2.89, \quad \sigma_3 = 2.58, \quad \sigma_4 = 2.79,$
$M_1 = 99.94, \quad M_2 = 73.54, \quad M_3 = 78.23, \quad M_4 = 77.39.$

Verify the following results:

$$\overline{X}_1 = 1093 - 2.09\,X_2 - 7.40\,X_3 - 3.36\,X_4,$$
$$\sigma_{1.234} = 21.69, \quad R_{1(234)} = .778.$$

4. The following regression equation was obtained by F. L. Whitney (*Journal of Educational Research*, May, 1923):

$$\overline{X}_1 = 23.218 + .004\,X_2 - .038\,X_3 - .115\,X_4 + .915\,X_5 + 1.403\,X_6 - .085\,X_7 \pm 3.02.$$

Predict the teaching success of a student with the following records:

$\overline{X}_1$ (to be predicted) = score on a rating scale.
$X_2 = 80$ = intelligence score,
$X_3 = 89.4$ = high-school academic record,
$X_4 = 8.7$ = normal-school academic record,
$X_5 = 8.5$ = normal-school professional record,
$X_6 = 8.6$ = student-teaching record,
$X_7 = 9.0$ = measure of physique.

($\overline{X}_1 = 38.2 \pm 3.02$. *Ans.*)

Interpret the regression coefficients. Do good academic work and good physique interfere with good teaching?

5. Derive the formula $R_{1(23)} = \sqrt{\dfrac{r_{12}^2 + r_{13}^2 - 2\,r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$.

6. Derive formulas (151a), (151b), and (151c).

7. Derive formulas (157), (158), and (159).

CHAPTER XVI

THE ELEMENTS OF CURVE-FITTING

1. Introductory

The investigator in many fields of science is frequently interested to determine the mathematical curve underlying his data. Such a curve is not only desirable in furnishing the theoretical law to which the observations conform, but is also of practical value as a basis for estimation. In the fields of education and psychology examples are furnished by learning curves, physical and mental growth curves, and frequency distributions. It is important to know the general laws of mental growth as well as to predict the standing of individuals of a given group, and for such purposes it is usually necessary to fit the data with a curve whose constants depend upon the observations. The plot of the experimental data often suggests some mathematical function which will be a good approximation to the observed material, allowing for the minor fluctuations in sampling. The problem is then to select the type of curve which is to be fitted to the data and to obtain its equation by appropriate methods. The suitability of the curve selected may finally be determined by tests for goodness of fit.

The choice of the proper sort of mathematical function will depend a great deal upon the worker's experience in curve-fitting and the accuracy of fit required. It is a well-known fact that by putting as many constants into the equation as there are observations the resulting curve will pass through all the observed points. If this is done, however, an extremely complicated function will result and the minor fluctuations, which should be smoothed out, will be given undue emphasis. It is therefore better to use a simple function involving only a

few constants, securing in this way a smoothing or graduation of the data which allows for the small fluctuations of sampling.

In the present chapter we shall introduce several of these simple curves and show how they may be fitted to the observed data. The observations to be fitted may consist of a series of points resulting from two measured characters such as the amount learned in a given time, or they may be given in the form of a frequency distribution. Three types of curves will be presented for fitting data of the first sort, while the normal probability curve will be used to illustrate the method of graduating frequency distributions. It should be noted that these curves have been selected from a very large number available because they have been found to give good results with certain data. They are presented here chiefly for illustration of the methods of fitting.

2. Types of Curves

In dealing with growth and learning data one of the most useful functions is the *hyperbola*, which for the purpose of curve-fitting may be most conveniently written in the form

$$Y = \frac{X}{a + bX} + c. \qquad \{\text{Hyperbola}\} \qquad (171)$$

The constants a, b, and c are to be determined from the observations. The use of this curve will be illustrated in applying the method of averages in section 5.

Another curve which has been found to give a good approximation to growth data is the *logarithmic growth function*,

$$Y = a + bX + c \log X. \quad \{\text{Logarithmic growth function}\} \quad (172)$$

This curve is similar in appearance to (171) and will be shown to give approximately as good results with certain data. The introduction of the terms $a + bX$ has the effect of raising and stretching out horizontally the ordinary logarithmic curve, $Y = c \log X$.

A third and very useful function is the *nth-order parabola,*

$$Y = C_0 + C_1X + C_2X^2 + C_3X^3 + \cdots + C_nX^n, \quad \{n\text{th-order parabola}\} \quad (173)$$

where the C's are constants determined by the data. If $C_2 = C_3 = \cdots = C_n = 0$, this expression reduces to the equation of a straight line; if $C_3 = C_4 = \cdots = C_n = 0$, an ordinary parabola results; while if $C_4 = C_5 = \cdots = C_n = 0$, a cubic is obtained, etc. In the case of regression curves from correlation tables, a very good fit is often obtained by the use of the nth-order parabola, but the question of how many terms to include must frequently be decided by trial and error.

A full discussion of frequency curves is beyond the scope of the present text. We shall therefore confine our illustration in the last section to the normal curve $y = y_0 e^{-\frac{x^2}{2\sigma^2}}$, which is already familiar to the reader.

3. Methods of Curve-Fitting

The first step in anticipation of curve-fitting is to plot the observed data so as to note the trend of the points and to determine, if possible, the appropriate curve to use. Having chosen some simple form such as described above, it is next necessary to determine the approximate values of the constants appearing in the equation. The methods used for such determination will depend upon the degree of accuracy required in the fit. If only a rough idea of the trend is required, a *free-hand curve* drawn through the observed points may be sufficient. For more accurate results, however, it will be necessary to apply certain mathematical methods known as *averages, least squares,* or *moments.* The first three of these methods will next be described, while the method of moments will be treated in sections 8 and 9 in dealing with frequency data.

Free-hand Method

A free-hand curve drawn through the observed points is clearly the easiest and simplest method to employ, but as already noted it may give results which are quite inaccurate. Several workers, moreover, would not agree closely upon the same free-hand graduation.

In drawing a curve through a series of points the fitting is often facilitated by the use of curved pieces of celluloid (French curves). These may be moved about as the curve is drawn so that the largest possible number of observed points lie on the curve or deviate equally on either side.

It sometimes happens that the most elaborate mathematical methods fail to give a good fit with certain data over a part of the range. In such cases it may be desirable to resort to free-hand approximations, possibly in combination with the other methods.*

Method of Averages

A second and more accurate method of curve-fitting is the method of averages. If Y represents an observed ordinate and $\overline{Y}$ denotes an ordinate on the fitted curve, the vertical deviations $Y - \overline{Y}$ are known as *residuals* (see Chapter IX). It is assumed in the method of averages that the "best" fit is that which makes the algebraic sum of the residuals equal to zero.

In the case of a straight line $\overline{Y} = a + bX$ the above condition requires that

$$\Sigma(Y - \overline{Y}) = \Sigma(Y - a - bX) = 0,$$

or

$$\Sigma Y - Na - b\Sigma X = 0. \qquad (174)$$

By dividing data into two parts, two equations of this type may be formed and solved for the constants a and b.

* For an example of this sort see an article by the author, "On the Relation of Vital Capacity to Certain Psychical Characters," *Biometrika*, Vol. XVI, p. 140.

While the method of averages may be used with functions involving several constants, it will be found most convenient when applied to the straight line where only two constants are required. This method will be illustrated in detail in section 5.

Method of Least Squares

The third and one of the best methods of curve-fitting is known as the method of least squares. This procedure, it will be recalled, has already been used in Chapter IX in obtaining the equations of the regression lines. Further illustration will now be given for the nth-order parabola.

If $\bar{Y}$ represents a value on such a parabola and Y an observation, the problem is to find values of the constants C_0, C_1, $C_2 \cdots C_n$ such that the sum of the squares of the residuals $(Y - \bar{Y})$ is as small as possible, that is, to have

$$u = \Sigma(Y - C_0 - C_1X - C_2X^2 - \cdots - C_nX^n)^2 = \text{a minimum.}$$

This is accomplished by equating to zero the partial derivatives* of u with respect to C_0, C_1, $C_2 \cdots C_n$ and thereby obtaining $n+1$ equations for the solution of the $n+1$ constants.

Differentiating in this way and setting the obtained results equal to zero, we find

$$\frac{\partial u}{\partial C_0} = 2\Sigma(Y - C_0 - C_1X - C_2X^2 - \cdots - C_nX^n)(-1) = 0,$$

$$\frac{\partial u}{\partial C_1} = 2\Sigma(Y - C_0 - C_1X - C_2X^2 - \cdots - C_nX^n)(-X) = 0,$$

$$\frac{\partial u}{\partial C_2} = 2\Sigma(Y - C_0 - C_1X - C_2X^2 - \cdots - C_nX^n)(-X^2) = 0,$$

. . . . . .

$$\frac{\partial u}{\partial C_n} = 2\Sigma(Y - C_0 - C_1X - C_2X^2 - \cdots - C_nX^n)(-X^n) = 0;$$

or, rearranging,

* See any good calculus, such as W. A. Granville, Differential and Integral Calculus. Ginn and Company.

$$\Sigma Y = C_0\Sigma(1) + C_1\Sigma(X) + C_2\Sigma(X^2) + \cdots + C_n\Sigma(X^n), \quad (175a)$$

$$\Sigma XY = C_0\Sigma(X) + C_1\Sigma(X^2) + C_2\Sigma(X^3) + \cdots + C_n\Sigma(X^{n+1}), \quad (175b)$$

$$\Sigma X^2Y = C_0\Sigma(X^2) + C_1\Sigma(X^3) + C_2\Sigma(X^4) + \cdots + C_n\Sigma(X^{n+2}), \quad (175c)$$

. . . . . .

$$\Sigma X^nY = C_0\Sigma(X^n) + C_1\Sigma(X^{n+1}) + C_2\Sigma(X^{n+2}) + \cdots + C_n\Sigma(X^{2n}), \quad (175d)$$

{Normal equations, unweighted ordinates}

where $\Sigma(1)$ = the number of ordinates summed.

These last expressions are known as the *normal equations*, of which there are clearly $n + 1$ in number. The variable Y is given by the observed ordinates taken from a convenient origin, while X may also be measured as the deviation $\cdots, -3, -2, -1, 0, 1, 2, 3, \cdots$ from any arbitrary point. The quantities ΣY, ΣXY, ΣX^2Y, etc., are found in the manner illustrated in section 7, and the $n + 1$ resulting linear equations are then solved for $C_0, C_1, C_2 \cdots C_n$.

In the above case it has been assumed that the ordinates Y are to be given equal weight. If these are obtained from the means of arrays, however, the frequencies f_x may need to be taken into account. It is then necessary to make the sum

$$\Sigma(f_xY - C_0f_x - C_1f_xX - C_2f_xX^2 - \cdots - C_nf_xX^n)^2 = \text{a minimum},$$

giving rise to the following normal equations:

$$\Sigma f_xY = C_0\Sigma(f_x) + C_1\Sigma(f_xX) + C_2\Sigma(f_xX^2) + \cdots + C_n\Sigma(f_xX^n), \quad (176a)$$

$$\Sigma f_xXY = C_0\Sigma(f_xX) + C_1\Sigma(f_xX^2) + C_2\Sigma(f_xX^3) + \cdots + C_n\Sigma(f_xX^{n+1}), \quad (176b)$$

. . . . . .

$$\Sigma f_xX^nY = C_0\Sigma(f_xX^n) + C_1\Sigma(f_xX^{n+1}) + C_2\Sigma(f_xX^{n+2}) + \cdots + C_n\Sigma(f_xX^{2n}). \quad (176c)$$

{Normal equations, weighted ordinates}

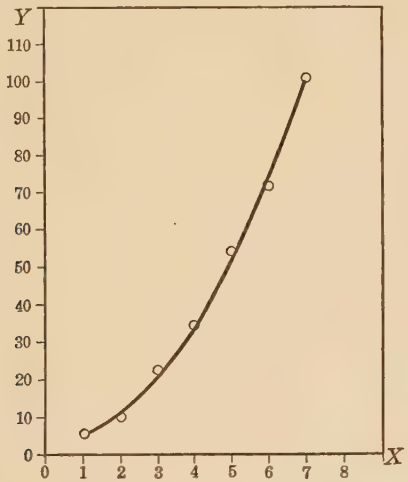
To distinguish these two methods the former is said to be based on *unweighted*, and the latter on *weighted, ordinates*. Both methods will be fully illustrated in section 7.

In comparing the fit of two or more curves to a given body of data, a good test is furnished by finding the squared or mean squared sum of the residuals, that is,

$$\Sigma(Y-\overline{Y})^2 \quad \text{or} \quad \frac{\Sigma(Y-\overline{Y})^2}{N}.$$

4. Illustration of the Free-hand Method

As an example of the free-hand method, we have selected a series of seven observations given at the right of Fig. 70. The curve was so drawn as to let the points deviate about equally on either side.



Y	X
5	1
10	2
22	3
35	4
55	5
73	6
101	7

Fig. 70. A free-hand curve drawn through seven observed points

If an approximation to the equation of such a curve is desired, it may often be found by *rectification*. Thus if the desired equation has the form

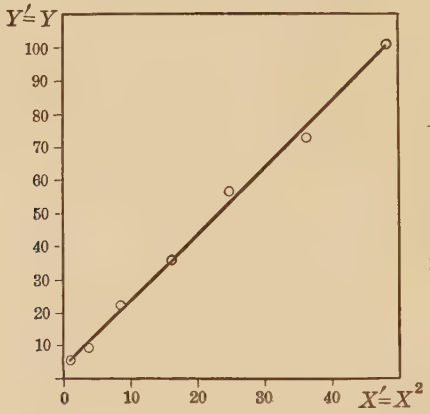
$$f(Y) = a + bF(X),^* \tag{177}$$

* The symbols $f(Y)$ and $F(X)$ mean a function of Y and a function of X. See Chapter III, section 5.

we may rectify this equation by substituting $Y' = f(Y)$ and $X' = F(X)$. The result,

$$Y' = a + bX', \tag{178}$$

will then be a straight line by means of which the constants a and b can be determined. The form of the original function



Y	X^2
5	1
10	4
22	9
35	16
55	25
73	36
101	49

FIG. 71. Illustrating the method of rectification

(177) must, of course, be guessed, but if a straight line results from (178) the choice is justified.

In the above problem it looks as if the desired equation might be a parabola of the form

$$Y = a + bX^2. \tag{179}$$

Setting $Y' = Y$ and $X' = X^2$, we may then find the plot of Y and X^2 to see if a straight line is obtained.

The graph in Fig. 71 clearly justifies the choice of the parabola, so that it only remains to obtain the constants a and b. Since the first and last points appear to fall on the line, we may obtain approximate values for these quantities by solving the resulting equations, $5 = a + b$ and $101 = a + 49b$, giving $a = 3$ and $b = 2$. The equation of the parabola is then

$$Y = 3 + 2X^2.$$

The method of rectification is useful not only in justifying the form of the function assumed but in furnishing a simple method of obtaining the necessary constants.

5. Fitting a Learning Curve with a Hyperbola by the Method of Averages

The data used for the present illustration were interpolated from a graph* showing the number of words typed in four minutes, Y, for various numbers of pages written, X. Inspection of Table 83 and Fig. 73, where the data are plotted, suggests that a hyperbola might be a good fit, and this is the curve employed by Thurstone.

Table 83. Data from L. L. Thurstone's Experiment in Typewriting

Total Number of Pages Written	Words Typed in Four Minutes (Average of 51 Subjects)
250	148
230	145
210	138
190	133
170	130
150	120
130	113
110	110
90	99
70	90
50	73
30	60
10	39

Inasmuch as the curve does not pass through the origin, it will be necessary to add a constant term to the equation of the hyperbola through (0, 0), with the result that

$$Y = \frac{X}{a + bX} + c. \tag{171}$$

* L. L. Thurstone, "The Learning Curve Equation," *Psychological Review Monographs*, Vol. XXV, No. 3 (1919), p. 45, Fig. 5. (Only the odd ordinates were used.)

The constant c here represents the skill in typewriting which the students have at the beginning of the experiment.

The above hyperbola may be *rectified* by selecting a point (X_k, Y_k) on the curve (or one that looks as if it might fall on the curve) and forming the differences

$$Y - Y_k = \frac{a(X - X_k)}{(a + bX)(a + bX_k)}.$$

The hyperbola passing through the point X_k, Y_k is therefore

$$\frac{X - X_k}{Y - Y_k} = (a + bX_k) + \frac{b}{a}(a + bX_k)X. \qquad (180)$$

Setting $\frac{X - X_k}{Y - Y_k} = Z$, $a + bX_k = m$, and $\frac{b}{a}(a + bX_k) = n$, we may also write

$$Z = m + nX, \qquad (181)$$

which is linear in X and Z. Thus if the plot of X and Z approximates a straight line, the original data will be approximated by the hyperbola (180), and the equation of the latter may be obtained from (181) by determining m and n.

TABLE 84. SHOWING THE CALCULATION NECESSARY FOR RECTIFYING THE HYPERBOLA FITTED TO THE LEARNING-CURVE DATA

PAGES	X	Y = WORDS IN 4 MINUTES	$X - 1$	$Y - 39$	$\frac{X-1}{Y-39} = Z$
250	13	148	12	109	.1101
230	12	145	11	106	.1038
210	11	138	10	99	.1010
190	10	133	9	94	.0957
170	9	130	8	91	.0879
150	8	120	7	81	.0864
130	7	113	6	74	.0811
110	6	110	5	71	.0704
90	5	99	4	60	.0667
70	4	90	3	51	.0588
50	3	73	2	34	.0588
30	2	60	1	21	.0476
10	1	39	0	0	—
Total	63				.5849
	28				.3834

The calculation for such rectification is shown in Table 84. The first point in the series with coördinates $X_k = 1$, $Y_k = 39$ has been selected for the origin. In Fig. 72 the values $Z = \dfrac{X-1}{Y-39}$ have been plotted with a resulting trend that appears to be fairly linear. It now remains to find the equation of the line of best average fit.

By dividing the data into two parts and summing over each, as shown in Table 84, the two equations like (181) necessary for the determination of m and n may be written

$$.5849 = 6\,m + 63\,n$$
and $$.3834 = 6\,m + 27\,n.$$

It will be noted that ΣX is reduced to 27 when only six items are used. Subtracting the second equation from the first to eliminate m, we find $n = .00560$ and, by substitution, $m = .0387$. The required straight line which is shown in Fig. 72 then has the equation

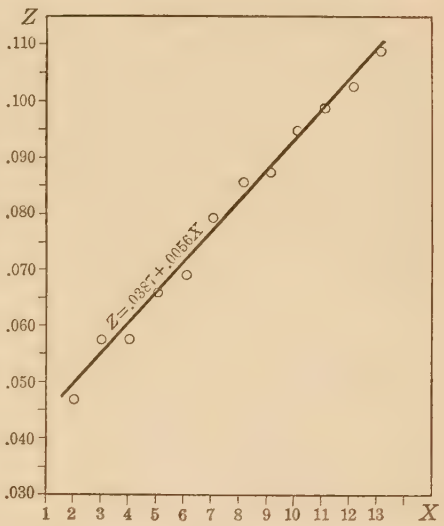


FIG. 72. Illustrating the method of rectifying the hyperbola for Thurstone's data

$$Z = .0387 + .0056\,X.$$

The equation for the hyperbola may now be written

$$\frac{X-1}{Y-39} = .0387 + .0056\,X,$$

or

$$Y = \frac{X-1}{.0387 + .0056\,X} + 39. \qquad (182)$$

A list of values for plotting equation (182) is given in Table 85.

TABLE 85. SHOWING THE CALCULATION OF THE ORDINATES FOR THE HYPERBOLA AND TEST FOR FIT BY SQUARED DIFFERENCES

X	.0056 X	.0056 X + .0387	$(X-1) \div$ (.0056 X +.0387)	$\overline{Y}$ = PREVIOUS COLUMN + 39	$Y - \overline{Y}$ (TABLE 84 FOR Y)	$(Y - \overline{Y})^2$
13	.0728	.1115	108	147	1	1
12	.0672	.1059	104	143	2	4
11	.0616	.1003	100	139	−1	1
10	.0560	.0947	95	134	−1	1
9	.0504	.0891	90	129	1	1
8	.0448	.0835	84	123	−3	9
7	.0392	.0779	77	116	−3	9
6	.0336	.0723	69	108	2	4
5	.0280	.0667	60	99	0	0
4	.0224	.0611	49	88	2	4
3	.0168	.0555	36	75	−2	4
2	.0112	.0499	20	59	1	1
1	.0056	.0443	0	39	0	0
						$\Sigma(Y - \overline{Y})^2 = 39$

From Fig. 73, where the hyperbola has been plotted, the fit appears to be a very good one. A numerical measure of fit is shown in the above table by the quantity

$$\Sigma(Y - \overline{Y})^2 = 39,$$

or

$$\frac{\Sigma(Y - \overline{Y})^2}{13} = 3.$$

This result will be compared later with that obtained in the case of a logarithmic growth curve. The size of

$$\Sigma(Y - \overline{Y})^2$$

will determine which curve is the better fit for the same observations. The χ^2 test is not feasible here, owing to difficulties which are beyond the scope of this text.

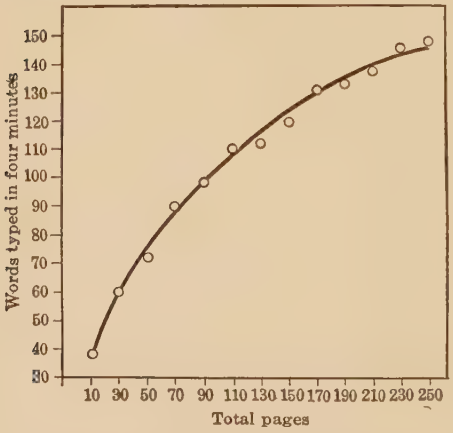


FIG. 73. Thurstone's data fitted with a hyperbola by the method of averages

6. Fitting a Learning Curve with the Logarithmic Growth Function by the Method of Least Squares

The data in the preceding section will next be fitted by the logarithmic growth curve

$$Y = a + bX + c \log X, \tag{172}$$

using the method of least squares with unweighted ordinates. The use of weighted ordinates is usually not necessary, and in the above problem the frequencies are not given.

It is now necessary to find the values of a, b, and c which will make the quantity

$$v = \Sigma(Y - a - bX - c \log X)^2 = \text{a minimum.}$$

The partial derivatives of v with respect to a, b, and c are next formed and equated to zero, as on page 321. The desired normal equations may then be written in the form

$$\Sigma(Y) = a\Sigma(1) + b\Sigma(X) + c\Sigma(\log X), \tag{183a}$$

$$\Sigma(XY) = a\Sigma(X) + b\Sigma(X^2) + c\Sigma(X \log X), \tag{183b}$$

$$\Sigma(Y \log X) = a\Sigma(\log X) + b\Sigma(X \log X) + c\Sigma(\log X)^2. \tag{183c}$$

{Normal equations for the logarithmic growth curve}

These may be solved for a, b, and c, giving the constants necessary for the logarithmic function of least-square fit.

The arithmetic is greatly facilitated by a table for sums such as $\Sigma(\log X)$, $\Sigma(X \log X)$, and $\Sigma(\log X)^2$, which is given on page 330. For a more extended table of these values the student should consult Pearl's "Medical Statistics," p. 368.

Upon examining equations (183) it is apparent that the quantities $\Sigma(X)$, $\Sigma(Y)$, $\Sigma(XY)$, $\Sigma(X^2)$, and $\Sigma(Y \log X)$ need to be calculated from the data, the remaining sums being obtained from Table 86. The calculation of these required sums is shown in full in Table 87, where, it will be noted, a check on $\Sigma(\log X)$ is obtained.

TABLE 86. SUMS OF LOG X, X LOG X, AND (LOG X)^2

X	Σ(Log X)	Σ(X Log X)	Σ(Log X)^2
1	0.00000	0.00000	0.00000
2	0.30103	0.60206	0.09062
3	0.77815	2.03342	0.31826
4	1.38021	4.44166	0.68074
5	2.07918	7.93651	1.16930
6	2.85733	12.60542	1.77482
7	3.70243	18.52111	2.48901
8	4.60552	25.74583	3.30458
9	5.55976	34.33401	4.21516
10	6.55976	44.33401	5.21516
11	7.60116	55.78933	6.29966
12	8.68034	68.73950	7.46429
13	9.79428	83.22077	8.70516
14	10.94041	99.26656	10.01877
15	12.11650	116.90793	11.40196
16	13.32062	136.17385	12.85187
17	14.55107	157.09148	14.36587
18	15.80634	179.68639	15.94158
19	17.08509	203.98270	17.57679
20	18.38612	230.00330	19.26947
21	19.70834	257.76991	21.01773
22	21.05077	287.30321	22.81983
23	22.41249	318.62295	24.67413
24	23.79271	351.74802	26.57912
25	25.19065	386.69652	28.53335

The calculation of $\Sigma(X)$ and $\Sigma(X^2)$ is facilitated by the use of Pearson's Tables XXVII and XXVIII, which give the sums and sums of powers of natural numbers.

The normal equations may now be written

$$(a) \qquad 1{,}398 = 13\,a + 91\,b + 9.7943\,c.$$

$$(b) \qquad 11{,}326 = 91\,a + 819\,b + 83.2208\,c.$$

$$(c) \qquad 1{,}187.8 = 9.7943\,a + 83.2208\,b + 8.7052\,c.$$

These may be solved by determinants, but straightforward elimination is probably as convenient as any method. The complete solution is given below for the benefit of those students who have not worked problems of this sort for some time. Multiplying equation (a) by 7 and subtracting from (b) gives

$$(d) \qquad 1540 = 182\,b + 14.6607\,c.$$

TABLE 87. SHOWING THE FORMATION OF SUMS NECESSARY FOR FITTING A LOGARITHMIC FUNCTION BY UNWEIGHTED ORDINATES

PAGES	X	Y = WORDS IN 4 MINUTES	XY	X^2	LOG X*	Y LOG X
250	13	148	1,924	169	1.11394	164.86312
230	12	145	1,740	144	1.07918	156.48110
210	11	138	1,518	121	1.04139	143.71182
190	10	133	1,330	100	1.00000	133.00000
170	9	130	1,170	81	0.95424	124.05120
150	8	120	960	64	0.90309	108.37080
130	7	113	791	49	0.84510	95.49630
110	6	110	660	36	0.77815	85.59650
90	5	99	495	25	0.69897	69.19803
70	4	90	360	16	0.60206	54.18540
50	3	73	219	9	0.47712	34.82976
30	2	60	120	4	0.30103	18.06180
10	1	39	39	1	0.00000	0.00000
Total . . .	91	1,398	11,326	819	9.79427	1,187.84583

Multiplying (a) by .75341 and subtracting from (c), we find

$$(e) \qquad 134.5 = 14.6605\, b + 1.3261\, c.$$

The terms involving b may next be eliminated by multiplying (e) by 12.4143 and combining with (d), with the result

$$(f) \qquad 129.7 = 1.8019\, c, \text{ or } c = 71.98.$$

By substitution and check we also obtain $a = 34.67$ and $b = 2.663$.

The required growth curve then has the equation

$$\overline{Y} = 34.67 + 2.663\, X + 71.98 \log X, \qquad (184)$$

and is plotted in Fig. 74.

From Table 88, where values for plotting are computed, it will also be noted that the sum of the squared differences, $\Sigma(Y - \overline{Y})^2$, is 48, which is not much larger than that obtained for the hyperbola fitted to the same data. In this example, then, there is little choice between the two curves.

* These values were read from an ordinary five-place logarithm table.

It should finally be noted that quite different forms of learning curves result when the time is recorded instead of the amount learned per unit of practice or time.* The data when

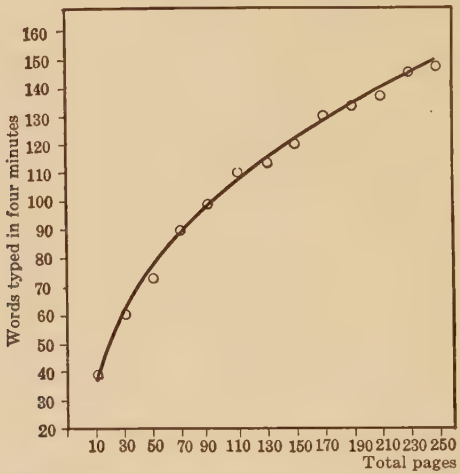


FIG. 74. Thurstone's data fitted by a logarithmic growth curve using the method of least squares

TABLE 88. VALUES FOR PLOTTING $\bar{Y} = 34.67 + 2.663\,X + 71.98\,\text{LOG}\,X$

PAGES	X	$2.663\,X$	$71.98\,\text{LOG}\,X$	ORDINATE, $\bar{Y}$	$(Y-\bar{Y})^2$
250	13	34.619	80.181	149.5	2.25
230	12	31.956	77.679	144.3	.49
210	11	29.293	74.959	138.9	.81
190	10	26.630	71.980	133.3	.09
170	9	23.967	68.686	127.3	7.29
150	8	21.304	65.004	121.0	1.00
130	7	18.641	60.830	114.1	1.21
110	6	15.978	56.011	106.7	10.89
90	5	13.315	50.312	98.3	.49
70	4	10.652	43.336	88.7	1.69
50	3	7.989	34.343	77.0	16.00
30	2	5.326	21.668	61.7	2.89
10	1	2.663	00.000	37.3	2.89
				$\Sigma(Y-\bar{Y})^2 = 47.99$	

* See Thurstone's Monograph cited above.

recorded in time units may be converted into amount per unit of time, as shown in Chapter VI, section 8, and then treated as illustrated above.

7. Fitting a Growth Curve with a Cubic by the Method of Least Squares

The following data were obtained from a correlation table of age and ossification ratio, the latter being the quotient of the ossified wrist-bone area divided by the area of a quadrilateral inclosing the carpal bones. The subjects were 520 boys in the Laboratory Schools of The University of Chicago. The measurements were made within a few days of each birthday.

Table 89. Data from Laboratory Schools

Central Age	Frequency	Mean Ossification Ratio
19	3	1.120
18	13	1.139
17	39	1.091
16	54	1.055
15	84	1.018
14	63	0.971
13	48	0.920
12	44	0.827
11	38	0.757
10	36	0.674
9	30	0.570
8	24	0.499
7	21	0.441
6	15	0.360
5	8	0.261
Total	520	

We shall fit a cubic to these data, first by considering the ordinates of equal weights, and then by weighted ordinates, using the observed frequencies as weights.

From equations (175), it is apparent that the quantities $\Sigma(Y) \cdots \Sigma(X^3Y)$, and $\Sigma(X) \cdots \Sigma(X^6)$ will be required. The

arithmetic is most easily done on a machine by the continuous process, that is, multiplying out the sub-products and adding them cumulatively on the calculator without separate listing. The complete work is shown, however, in the accompanying table. It will be noted that X has been measured from the central age, 12, which makes the sums of the odd powers of X equal to zero.

TABLE 90. SHOWING THE FORMATION OF SUMS NECESSARY FOR FITTING A CUBIC BY THE METHOD OF UNWEIGHTED ORDINATES

X AGE — 12	Y	XY	X^2Y	X^3Y	X^2	X^3	X^4	X^5	X^6
7 . .	1.120	7.840	54.880	384.160	49	343	2,401	16,807	117,649
6 . .	1.139	6.834	41.004	246.024	36	216	1,296	7,776	46,656
5 . .	1.091	5.455	27.275	136.375	25	125	625	3,125	15,625
4 . .	1.055	4.220	16.880	67.520	16	64	256	1,024	4,096
3 . .	1.018	3.054	9.162	27.486	9	27	81	243	729
2 . .	0.971	1.942	3.884	7.768	4	8	16	32	64
1 . .	0.920	.920	.920	.920	1	1	1	1	1
0 . .	0.827	—	—	—	—	—	—	—	—
− 1 . .	0.757	− .757	.757	− .757	1	− 1	1	− 1	1
− 2 . .	0.674	− 1.348	2.696	− 5.392	4	− 8	16	− 32	64
− 3 . .	0.570	− 1.710	5.130	− 15.390	9	− 27	81	− 243	729
− 4 . .	0.499	− 1.996	7.984	− 31.936	16	− 64	256	− 1,024	4,096
− 5 . .	0.441	− 2.205	11.025	− 55.125	25	− 125	625	− 3,125	15,625
− 6 . .	0.360	− 2.160	12.960	− 77.760	36	− 216	1,296	− 7,776	46,656
− 7 . .	0.261	− 1.827	12.789	− 89.523	49	− 343	2,401	− 16,807	117,649
0	11.703	18.262	207.346	594.370	280	0	9,352	0	369,640

Equations (175) may now be written

$$(a) \qquad 11.703 = 15\,C_0 + 280\,C_2.$$

$$(b) \qquad 18.262 = 280\,C_1 + 9352\,C_3.$$

$$(c) \qquad 207.346 = 280\,C_0 + 9352\,C_2.$$

$$(d) \qquad 594.370 = 9352\,C_1 + 369{,}640\,C_3.$$

These may be solved by elimination, as illustrated in the preceding section, giving

$$C_0 = +\,.8305,\ C_1 = +\,.0743,\ C_2 = -\,.002693,$$

and

$$C_3 = -\,.000272.$$

The required cubic is therefore

$$\bar{Y} = .8305 + .0743X - .002693X^2 - .000272X^3. \qquad (185)$$

In order to compare results with those obtained by the following method, the origin will be shifted to age 13. Taken from this point, the equation becomes

$$\bar{Y} = .8305 + .0743(X_1+1) - .002693(X_1+1)^2 - .000272(X_1+1)^3$$

or
$$\bar{Y} = .902 + .0681 X_1 - .00351 X_1^2 - .000272 X_1^3. \qquad (186)$$

Before plotting this result together with the observed points, the equation of the cubic by the method of weighted ordinates will next be obtained. The arithmetic is much lengthier because none of the terms vanish as above. Table 91 shows the full calculation for the sums entering into equations (176), each of the totals being divided by 520 to give more convenient numbers.

Forming equations (176), we find

(a) $.8534 = C_0 - .2519\ C_1 + 10.663\ C_2 - 25.106\ C_3.$
(b) $.5047 = -.2519\ C_0 + 10.663\ C_1 - 25.106\ C_2 + 288.51\ C_3.$
(c) $7.2352 = 10.663\ C_0 - 25.106\ C_1 + 288.51\ C_2 - 1295.2\ C_3.$
(d) $-1.6374 = -25.106\ C_0 + 288.51\ C_1 - 1295.2\ C_2 + 11{,}377\ C_3.$

The elimination is next given in detail for illustration.
Multiplying (a) by .2519 and adding to (b) gives

$$(e)\ .7197 = 10.600\ C_1 - 22.42\ C_2 + 282.19\ C_3.$$

Multiplying (a) by 10.663 and subtracting (c), we obtain

$$(f)\ 1.8646 = 22.42\ C_1 - 174.81\ C_2 + 1027.5\ C_3,$$

and multiplying (a) by 25.106 and adding to (d), we find

$$(g)\ 19.788 = 282.19\ C_1 - 1027.5\ C_2 + 10{,}747\ C_3.$$

This gives three equations in three unknowns.
Terms in C_1 are next eliminated as follows:

(h) $.628 = -430.6\ C_2 + 3235\ C_3$ $[(g) - 26.622 \times (e)]$.
(i) $.1619 = -60.23\ C_2 + 203.61\ C_3$ $[.4728 \times (f) - (e)]$.

TABLE 91. SHOWING THE FORMATION OF THE SUMS NECESSARY FOR FITTING A CUBIC BY THE METHOD OF WEIGHTED ORDINATES

X AGE — 13	Y	f	fY	fXY	fX^2Y	fX^3Y	fX	fX^2	fX^3	fX^4	fX^5	fX^6
6 . .	1.120	3	3.360	20.160	120.960	725.760	18	108	648	3,888	23,328	139,968
5 . .	1.139	13	14.807	74.035	370.175	1,850.875	65	325	1,625	8,125	40,625	203,125
4 . .	1.091	39	42.549	170.196	680.784	2,723.136	156	624	2,496	9,984	39,936	159,744
3 . .	1.055	54	56.970	170.910	512.730	1,538.190	162	486	1,458	4,374	13,122	39,366
2 . .	1.018	84	85.512	171.024	342.048	684.096	168	336	672	1,344	2,688	5,376
1 . .	.971	63	61.173	61.173	61.173	61.173	63	63	63	63	63	63
0 . .	.920	48	44.160	—	—	—	—	—	—	—	—	—
− 1 . .	.827	44	36.388	− 36.388	36.388	− 36.388	− 44	44	− 44	44	− 44	44
− 2 . .	.757	38	28.766	− 57.532	115.064	− 230.128	− 76	152	− 304	608	− 1,216	2,432
− 3 . .	.674	36	24.264	− 72.792	218.376	− 655.128	− 108	324	− 972	2,916	− 8,748	26,244
− 4 . .	.570	30	17.100	− 68.400	273.600	− 1,094.400	− 120	480	− 1,920	7,680	− 30,720	122,880
− 5 . .	.499	24	11.976	− 59.880	299.400	− 1,497.000	− 120	600	− 3,000	15,000	− 75,000	375,000
− 6 . .	.441	21	9.261	− 55.566	333.396	− 2,000.376	− 126	756	− 4,536	27,216	− 163,296	979,776
− 7 . .	.360	15	5.400	− 37.800	264.600	− 1,852.200	− 105	735	− 5,145	36,015	− 252,105	1,764,735
− 8 . .	.261	8	2.088	− 16.704	133.632	− 1,069.056	− 64	512	− 4,096	32,768	− 262,144	2,097,152
Total		520	443.774	262.436	3,762.326	− 851.446	− 131	5,545	−13,055	150,025	− 673,511	5,915,905
		1.0000	.8534	.5047	7.2352	− 1.6374	− .2519	10.663	− 25.106	288.51	− 1295.2	11,377

Multiplying (h) by .13987 and subtracting from (i) gives, finally, (j) $.0741 = -248.9\,C_3$. $\therefore\ C_3 = -.000298$.
Substituting this value in (h), $C_2 = -.00370$.
From (g) and (a), we also find $C_1 = .0680$ and $C_0 = .9025$.
The required cubic is therefore

$$\bar{Y} = .9025 + .0680\,X - .00370\,X^2 - .000298\,X^3. \qquad (187)$$

It will be noted that the coefficients in equations (186) and (187) are in close agreement except for the last two, where no great effect will be produced except for high values of X. Comparison of the two cubics is shown in Table 92, where values of $\bar{Y}$ have been tabulated with X taken from the origin $X = 13$. The plot of these results in Fig. 75 shows that the only noticeable difference in fit occurs for the high values of ossification ratio, but the number of cases in this range is so small that no very accurate smoothing is to be expected. Experience generally shows that the method of unweighted ordinates gives approximately as good results as the method of weighted ordinates, except when the weighting is very uneven.

TABLE 92. VALUES FOR PLOTTING CUBICS (186) AND (187)

AGE	X	ORDINATES $\bar{Y}$	
		For (186)	For (187)
20	7	1.113	1.095
19	6	1.125	1.113
18	5	1.121	1.113
17	4	1.101	1.096
16	3	1.067	1.065
15	2	1.022	1.021
14	1	0.966	0.966
13	0	0.902	0.902
12	−1	0.831	0.831
11	−2	0.754	0.754
10	−3	0.673	0.673
9	−4	0.591	0.590
8	−5	0.508	0.507
7	−6	0.426	0.426
6	−7	0.347	0.347
5	−8	0.272	0.274
4	−9	0.203	0.208

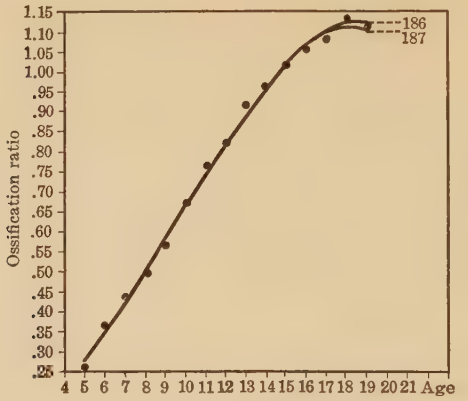


FIG. 75. Plot of the cubics (186) and (187)

8. THE METHOD OF MOMENTS APPLIED TO FREQUENCY DATA

In fitting frequency distributions with mathematical curves, one of the best and most widely used procedures is the *method of moments*, developed for this purpose by Professor Pearson. The graduation, or fit, is obtained by equating the moments of the data to the moments of the curve to be fitted.

If a frequency distribution be given with frequencies f_1, f_2, $f_3, \cdots f_t$, occurring at class values X_1, X_2, $X_3, \cdots X_t$, then the sum $f_1X_1 + f_2X_2 + f_3X_3 + \cdots + f_tX_t$ is called the *first moment* with reference to the origin from which X is measured. Similarly, $f_1X_1^2 + f_2X_2^2 + f_3X_3^2 + \cdots + f_tX_t^2$ is called the *second moment*, and $f_1X_1^3 + f_2X_2^3 + f_3X_3^3 + \cdots + f_tX_t^3$ is known as the *third moment*, etc. These quantities may be more briefly written as ΣfX, ΣfX^2, $\Sigma fX^3 \cdots$, so that the

$$p\text{th moment about the origin} = \Sigma fX^p. \qquad (188)$$

When each of the above moments has been divided by N, the result,

$$\bar{v}_p = \frac{\Sigma fX^p}{N}, \quad \left\{\begin{matrix}\text{Moment coefficient} \\ \text{about the origin}\end{matrix}\right\} \qquad (189)$$

has been termed by Professor Pearson a *moment coefficient* about the origin.

The moment coefficients about the mean are given by the formula

$$\nu_p = \frac{\Sigma f x^p}{N} = \frac{\Sigma f (X - M)^p}{N}. \quad \left\{\begin{array}{l}\text{Formula for moment coefficients}\\ \text{about the mean}\end{array}\right\} \quad (190)$$

The reader will note that the moment coefficient about the origin is denoted by $\bar{\nu}_p$, while the moment coefficient about the mean is given by ν_p. Substituting various values for p in (190), and observing that $\Sigma f = N$, we may write

$$\nu_0 = \frac{\Sigma f x^0}{N} = 1, \quad (191a)$$

$$\nu_1 = \frac{\Sigma f x}{N} = 0, \quad (191b)$$

$$\nu_2 = \frac{\Sigma f x^2}{N} = \sigma_x^2, \text{ etc.} \quad (191c)$$

{Moment coefficients about the mean}

Certain relationships between the moments about the origin and the mean may be obtained by expanding (190). Thus,

$$\nu_p = \frac{\Sigma f}{N}\left[X^p - pX^{p-1}M + \frac{p(p-1)}{1\cdot 2}X^{p-2}M^2 - \frac{p(p-1)(p-2)}{1\cdot 2\cdot 3}X^{p-3}M^3 + \cdots\right]$$

or,

$$\nu_p = \bar{\nu}_p - p\bar{\nu}_{p-1}\bar{\nu}_1 + \frac{p(p-1)}{2}\bar{\nu}_{p-2}\bar{\nu}_1^2 - \frac{p(p-1)(p-2)}{6}\bar{\nu}_{p-3}\bar{\nu}_1^3 + \cdots. \quad (192)$$

Since $\nu_0 = \bar{\nu}_0 = 1$, we find, upon setting $p = 1$, 2, 3, and 4 in this last equation, that

$$\nu_1 = \bar{\nu}_1 - \bar{\nu}_1 = 0, \quad (193a)$$

$$\nu_2 = \bar{\nu}_2 - \bar{\nu}_1^2, \quad (193b)$$

$$\nu_3 = \bar{\nu}_3 - 3\bar{\nu}_1\bar{\nu}_2 + 2\bar{\nu}_1^3, \quad (193c)$$

and that

$$\nu_4 = \bar{\nu}_4 - 4\bar{\nu}_1\bar{\nu}_3 + 6\bar{\nu}_1^2\bar{\nu}_2 - 3\bar{\nu}_1^4. \quad (193d)$$

{Moment coefficients about the mean in terms of those about the origin}

By taking the moments about the origin, we may also write

$$\bar{\nu}_p = \frac{\Sigma f X^p}{N} = \frac{\Sigma f(x+\bar{\nu}_1)^p}{N},$$

or

$$\bar{\nu}_p = \nu_p + p\nu_{p-1}\bar{\nu}_1 + \frac{p(p-1)}{2}\nu_{p-2}\bar{\nu}_1^2 + \frac{p(p-1)(p-2)}{6}\nu_{p-3}\bar{\nu}_1^3 + \cdots.$$

Transposing we then have

$$\nu_p = \bar{\nu}_p - p\nu_{p-1}\bar{\nu}_1 - \frac{p(p-1)}{2}\nu_{p-2}\bar{\nu}_1^2 - \frac{p(p-1)(p-2)}{6}\nu_{p-3}\bar{\nu}_1^3 - \cdots. \qquad (194)$$

Substituting values of p from 0 to 4 gives the following set of equations, which may be used as a check on equations (193):

$$\nu_0 = 1, \qquad (195a)$$

$$\nu_1 = 0, \qquad (195b)$$

$$\nu_2 = \bar{\nu}_2 - \bar{\nu}_1^2, \qquad (195c)$$

$$\nu_3 = \bar{\nu}_3 - 3\bar{\nu}_1\nu_2 - \bar{\nu}_1^3, \qquad (195d)$$

$$\nu_4 = \bar{\nu}_4 - 4\bar{\nu}_1\nu_3 - 6\bar{\nu}_1^2\nu_2 - \bar{\nu}_1^4. \qquad (195e)$$

{ Moment coefficients about the mean in terms of moments about the origin and mean }

The fifth, sixth, and higher moments might be formed in a similar way, but Professor Pearson* has shown that, except for very large samples, their probable errors are too high for the results to be of any value in curve-fitting.

It should be noted that equations (193) and (195) hold when X is measured from any origin, since $x = X - M = X' - M'$, where $X' = X - A$, A being the arbitrary origin. The moment coefficients about the mean may therefore be obtained by choosing an arbitrary point and making subsequent adjustment as in the case of the standard deviation.

* Karl Pearson, "Skew Correlation and Non-Linear Regression," *Draper's Research Memoirs II*. Cambridge University Press, 1905.

It is now necessary to distinguish two types of series which may arise:

a. The data may consist of a system of isolated ordinates as in the case of the point binomial. This type, however, will not be considered in the present treatment.

b. The data may consist of a system of areas as in the frequency distribution of a measured variable. Here the moments are calculated by assuming that the areas are concentrated at the class values and corrections for equations (191) to (195) are therefore necessary. These adjustments, which are known as Sheppard's * corrections, will next be given and the complete arithmetic shown for a distribution resembling the normal curve.

Denoting the moment coefficients adjusted for grouping by μ_1, μ_2, μ_3, and μ_4, Sheppard's correction may be written

$$\mu_1 = \nu_1, \tag{196a}$$

$$\mu_2 = \nu_2 - \tfrac{1}{12} = \nu_2 - .083333,\dagger \tag{196b}$$

$$\mu_3 = \nu_3, \tag{196c}$$

$$\mu_4 = \nu_4 - \tfrac{1}{2}\nu_2 + \tfrac{7}{240} = \nu_4 - .5\,\nu_2 + .02916667. \tag{196d}$$

{ Moment coefficients about the mean adjusted for grouping (Sheppard's corrections) }

The proof of these equations is based on the assumption that the derivatives of the frequency function vanish at the limits of the curve. The corrections are to be used therefore when the distribution has "high contact" at the extremes of the scale, that is, tapers off gradually at both ends.

Professor Karl Pearson has developed a number of curves for the purpose of describing biometric data. These curves, which vary from extremely skewed to symmetrical types, are identified by certain criteria worked out from the distributions to which

* W. F. Sheppard, Proceedings of the London Mathematical Society, Vol. XXIX, pp. 353–380.

† Note that $\sigma = \sqrt{\mu_2}\,h = (\sqrt{\nu_2 - \tfrac{1}{12}})h$.

the curves are to be fitted. Some of the constants used by Professor Pearson may be set down as follows:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \qquad \kappa_1 = 2\beta_2 - 3\beta_1 - 6,$$
$$\beta_2 = \frac{\mu_4}{\mu_2^2}, \qquad \kappa_2 = \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)}. \tag{197}$$

{Pearson's constants for curve-fitting}

It will be noted that β_1 and β_2 are independent of the units of measure of the distributed variables.

The steps in curve-fitting are then briefly as follows:

1. Work out the first four adjusted moment coefficients,

$$\mu_1, \mu_2, \mu_3, \text{ and } \mu_4.$$

2. Form β_1, β_2, κ_1, and κ_2, in order to determine which type of curve to employ.

3. Find the constants of the curve selected from the moments and the β's (formulas for the maximum ordinate and other parameters are given in Elderton* for each type of curve).

4. Plot the curve with a histogram of the data and note the general goodness of fit.

5. Test the goodness of fit by the χ^2 method, finding the areas under the curve by arithmetical or mechanical integration.

In the following section these steps will be illustrated by the normal probability curve.

9. Fitting a Normal Curve by the Method of Moments

The data selected for graduation consist of the heights of men in the British Isles (see Table 41, p. 206). These have been chosen because they furnish a fairly good example of normally distributed data and illustrate the simplest of Pearson's types of frequency curves.

* See Elderton's "Frequency Curves and Correlation," Jones's "First Course in Statistics," and Pearson's Tables, Introduction, for detailed discussion of these types of curves.

The criteria for the normal curve $y = y_0 e^{-\frac{x^2}{c}}$, which should be satisfied if this curve is appropriate, are

$$\beta_1 = 0 \tag{198}$$

and

$$\beta_2 = 3, \tag{199}$$

while the constants are determined by

$$c = 2\mu_2, \tag{200}$$

$$y_0 = \frac{N}{\sqrt{2\pi\mu_2}}, \tag{201}$$

and

$$M = 0 = \text{origin}. \tag{202}$$

{Criteria and constants for a normal curve}

It is now necessary to work out these values from the data, and compare with those given by equations (198) and (199). The constants for the curve are furnished by equations (200), (201), and (202).

In calculating the unadjusted moments the arithmetic may be conveniently arranged as illustrated by Table 93 on page 344. Using equation (189), we find from the values at the bottom of the table that $\bar{\nu}_1 = .020850$, $\bar{\nu}_2 = 6.617239$, $\bar{\nu}_3 = 0.206057$, and $\bar{\nu}_4 = 137.689109$. Substituting these values in equations (193) or (195), the unadjusted moment coefficients about the mean become

$$\nu_1 = 0,\ \nu_2 = 6.616804,\ \nu_3 = -.207833, \text{ and } \nu_4 = 137.689183.$$

The adjusted moment coefficients may now be found from equations (196), giving

$$\mu_1 = 0,\ \mu_2 = 6.533471,\ \mu_3 = -.207833, \text{ and } \mu_4 = 134.4099.$$

By substituting these last values in equations (197), where the general expressions for β_1 and β_2 are given, we find

$$\beta_1 = .000155$$

and

$$\beta_2 = 3.14879.$$

The values for κ_1 and κ_2 are not required in fitting the normal probability curve.

TABLE 93. SHOWING CALCULATION OF THE FIRST FOUR UNADJUSTED MOMENTS OF A FREQUENCY DISTRIBUTION

CENTRAL HEIGHT	f	d	fd	fd^2	fd^3	fd^4
$77\frac{7}{16}$	2	10	20	200	2,000	20,000
$76\frac{7}{16}$	5	9	45	405	3,645	32,805
$75\frac{7}{16}$	16	8	128	1,024	8,192	65,536
$74\frac{7}{16}$	32	7	224	1,568	10,976	76,832
$73\frac{7}{16}$	79	6	474	2,844	17,064	102,384
$72\frac{7}{16}$	202	5	1,010	5,050	25,250	126,250
$71\frac{7}{16}$	392	4	1,568	6,272	25,088	100,352
$70\frac{7}{16}$	646	3	1,938	5,814	17,442	52,326
$69\frac{7}{16}$	1,063	2	2,126	4,252	8,504	17,008
$68\frac{7}{16}$	1,230	1	1,230	1,230	1,230	1,230
$67\frac{7}{16}$	1,329	0	—	—	—	—
$66\frac{7}{16}$	1,223	−1	−1,223	1,223	−1,223	1,223
$65\frac{7}{16}$	990	−2	−1,980	3,960	−7,920	15,840
$64\frac{7}{16}$	669	−3	−2,007	6,021	−18,063	54,189
$63\frac{7}{16}$	394	−4	−1,576	6,304	−25,216	100,864
$62\frac{7}{16}$	169	−5	−845	4,225	−21,125	105,625
$61\frac{7}{16}$	83	−6	−498	2,988	−17,928	107,568
$60\frac{7}{16}$	41	−7	−287	2,009	−14,063	98,441
$59\frac{7}{16}$	14	−8	−112	896	−7,168	57,344
$58\frac{7}{16}$	4	−9	−36	324	−2,916	26,244
$57\frac{7}{16}$	2	−10	−20	200	−2,000	20,000
Totals	8,585		+179	56,809	+1,769	1,182,061
Unadjusted moments			$= N\bar{\nu}_1$	$= N\bar{\nu}_2$	$= N\bar{\nu}_3$	$= N\bar{\nu}_4$

The probable errors of β_1 and β_2 for samples from a normal population are given approximately by

$$P.E.\text{ of }\sqrt{\beta_1} \doteq \frac{.6745\sqrt{6}}{\sqrt{N}} \tag{203}$$

and

$$P.E.\text{ of }\beta_2 \doteq \frac{.6745\sqrt{24}}{\sqrt{N}}. \tag{204}$$

We may therefore write $\sqrt{\beta_1}=.012\pm.018$ and $\beta_2=3.149\pm.036$, and conclude that the normal curve is appropriate even though a value of β_2 as high as 3.149 is rather improbable.

When the goodness of fit is tested by χ^2 as in section 7 of Chapter XIII, it is found that the fit is satisfactory. This is left as an exercise for the student.

EXERCISES

1. Fit a hyperbola by the method of averages to the data in the accompanying table. Use the scale 1, 2, 3 · · · for pages written, and select $X_k = 1$, $Y_k = 30$ for rectifying point.

Pages Written	Words Typed in Four Minutes
370	192
350	188
330	184
310	172
290	195
270	178
250	180
230	164
210	161
190	160
170	151
150	142
130	137
110	122
90	106
70	100
50	81
30	57
10	30

$$\left(\overline{Y} = \frac{X - 1}{.027 + .0044\,X} + 30.\ \textit{Ans.}\right)$$

2. Fit the data of Exercise 1 with a logarithmic growth curve, using the method of least squares. Compare the fit with that obtained for the hyperbola.

$$(\overline{Y} = 22.56 + .526\,X + 127.1 \log X.\ \textit{Ans.})$$

3. The data on page 346 are the ossification ratios of 540 girls of the Laboratory Schools of The University of Chicago. Fit a cubic to the means by the method of least squares. (Use unweighted ordinates and take the origin at age 12.)

4. Calculate and plot the means of the columns from the table on page 189. Fit a cubic to these points by the method of least squares, using unweighted ordinates. Compare the equation with the following, based on more data:*

$$\overline{\beta} = 23.14 + 1.2545\,\alpha - .0089\,\alpha^2 + .000025\,\alpha^3.$$

* See Memoirs of the National Academy of Sciences, Vol. XV, p. 576.

Central Age	f	Mean Ossification Ratio
19	5	1.160
18	14	1.102
17	53	1.098
16	63	1.108
15	69	1.089
14	63	1.061
13	40	1.033
12	44	.988
11	38	.898
10	38	.834
9	39	.730
8	26	.662
7	17	.523
6	23	.442
5	8	.358
	540	

($y = .961 + .0576\,x - .00475\,x^2 - .0000230\,x^3$. *Ans.*)

5. Data: cephalic index of 1982 boys aged 13 (from Professor Pearson's laboratory).

Index	f	Index	f
91	1	78	293.5
90	1	77	236.5
89	4	76	181.5
88	4	75	156.5
87	7	74	78
86	23	73	49
85	31	72	23
84	58	71	26
83	93	70	8
82	130	69	8
81	156	68	2
80	181.5	67	3
79	227.5	Total	1982

Find μ_2, μ_3, μ_4, β_1, and β_2, and fit with a normal curve. Work out the chi-square test for goodness of fit.

($\mu_2 = 10.980$; $\mu_3 = 2.326$; $\mu_4 = 409.112$; $\beta_1 = .0041$; $\beta_2 = 3.393$; $y_0 = 238.62$; $P = .0001$, throwing together the five highest and also the four lowest groups. *Ans.*)

APPENDIX A

LIST OF IMPORTANT FORMULAS FOR REFERENCE*

$$M = \frac{\Sigma X}{N}. \qquad \left\{\begin{array}{c}\text{Mean for}\\ \text{ungrouped series}\end{array}\right\} \qquad (5)$$

$$M = \frac{\Sigma X}{N} = A + \left(\frac{\Sigma fd}{N}\right)h. \qquad \left\{\begin{array}{c}\text{Mean for}\\ \text{distribution}\end{array}\right\} \qquad (6)$$

$$Md = l.\,l. + \left(\frac{\frac{N}{2} - f_{up}}{f_{md}}\right)h, \qquad \left\{\begin{array}{c}\text{Median for}\\ \text{distribution}\\ \text{counting up}\end{array}\right\} \qquad (8a)$$

$$Md = u.\,l. - \left(\frac{\frac{N}{2} - f_{do}}{f_{md}}\right)h. \qquad \left\{\begin{array}{c}\text{Median for dis-}\\ \text{tribution count-}\\ \text{ing down}\end{array}\right\} \qquad (8b)$$

$$G.M. = \sqrt[N]{X_1\ X_2\ X_3 \cdot \cdot \cdot\ X_N}. \qquad \left\{\begin{array}{c}\text{Geometric}\\ \text{mean}\end{array}\right\} \qquad (9)$$

$$\log (G.M.) = \frac{1}{N}\Sigma \log (X). \qquad \left\{\begin{array}{c}\text{Logarithmic}\\ \text{form of geo-}\\ \text{metric mean}\end{array}\right\} \qquad (10)$$

$$\frac{1}{H} = \frac{1}{N}\Sigma\left(\frac{1}{X}\right). \qquad \{\text{Harmonic mean}\} \qquad (11)$$

$$M.D. = \frac{\Sigma |x|}{N}. \qquad \{\text{Mean deviation}\} \qquad (12)$$

$$M.D. = \frac{(\Sigma |fd|)h + (A_m - M)(N_a - N_b)}{N}. \qquad \left\{\begin{array}{c}\text{Mean deviation}\\ \text{for frequency}\\ \text{distribution}\end{array}\right\} \qquad (14)$$

$$S.D. = \sqrt{\frac{\Sigma x^2}{N}}. \qquad \left\{\begin{array}{c}\text{Standard deviation,}\\ \text{original form}\end{array}\right\} \qquad (15)$$

$$S.D. = \sqrt{\frac{\Sigma (X')^2}{N} - (M')^2}. \qquad \left\{\begin{array}{c}\text{Standard devia-}\\ \text{tion for reduced}\\ \text{series}\end{array}\right\} \qquad (16)$$

* For notation see list of important symbols in Appendix B.

$$S.D. = \left(\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}\right)h. \qquad \left\{\begin{matrix}\text{Standard de-}\\ \text{viation for}\\ \text{distribution}\end{matrix}\right\} \qquad (17)$$

$$Q = \frac{Q_3 - Q_1}{2}. \qquad \{\text{Quartile deviation}\} \qquad (19)$$

$$Q_3 = u.l. - \frac{\frac{N}{4} - f_{do}}{f_3} \times h, \qquad (20a)$$

$$\left\{\begin{matrix}\text{Quartiles}\\ \text{for distribution}\end{matrix}\right\}$$

$$Q_1 = l.l. + \frac{\frac{N}{4} - f_{up}}{f_1} \times h. \qquad (20b)$$

$$V = \frac{100\,\sigma}{M}. \qquad \{\text{Coefficient of variation}\} \qquad (22)$$

$$X_1 = M_1 + \frac{\sigma_1}{\sigma_2}(X_2 - M_2). \qquad \left\{\begin{matrix}\text{Transmutation formula}\\ \text{for comparable scores,}\\ \text{score form}\end{matrix}\right\} \qquad (23)$$

$$S_k = \frac{(Q_3 - Md) - (Md - Q_1)}{Q} = \frac{Q_1 + Q_3 - 2\,Md}{Q}. \qquad \left\{\begin{matrix}\text{Measure of}\\ \text{skewness}\\ \text{based on}\\ \text{quartiles}\end{matrix}\right\} \qquad (24)$$

$$S_k = \frac{M - M_o}{\sigma}. \qquad \left\{\begin{matrix}\text{Pearson's measure}\\ \text{of skewness}\end{matrix}\right\} \qquad (25)$$

$$S_k = \frac{3\,(M - Md)}{\sigma}. \qquad \left\{\begin{matrix}\text{Approximate meas-}\\ \text{ure of skewness}\end{matrix}\right\} \qquad (26)$$

$$P_p = l.l. + \left[\frac{\frac{pN}{100} - f_{up}}{f_P}\right]h, \qquad \left\{\begin{matrix}\text{Percentiles,}\\ \text{counting up}\end{matrix}\right\} \qquad (27a)$$

$$P_p = u.l. - \left[\frac{\left(\frac{100 - p}{100}\right)N - f_{do}}{f_P}\right]h. \qquad \left\{\begin{matrix}\text{Counting}\\ \text{down}\end{matrix}\right\} \qquad (27b)$$

$$R_x = R_l + \frac{R_u - R_l}{h}(X - l.l.). \qquad \left\{\begin{matrix}\text{Percentile}\\ \text{rank formula,}\\ \text{form 1}\end{matrix}\right\} \qquad (28)$$

$$R_x = \frac{100\,[f_x\,(X - l.l.) + (f_{up})h]}{Nh}. \qquad \left\{\begin{matrix}\text{Percentile rank}\\ \text{formula, form 2}\end{matrix}\right\} \qquad (29)$$

$$ {}_cR_x = \frac{50 f_x}{N} + \frac{100\,(f_{up})}{N} = \frac{50 f_x}{N} + R_l. \qquad \left\{\begin{matrix}\text{Class value}\\ \text{rank}\end{matrix}\right\} \qquad (30) $$

$$ r = \frac{\Sigma xy}{N\sigma_x\sigma_y}. \qquad \left\{\begin{matrix}\text{Product-moment}\\ \text{correlation coefficient,}\\ \text{original form}\end{matrix}\right\} \qquad (31) $$

$$ r = \frac{\Sigma xy}{\sqrt{\Sigma x^2\ \Sigma y^2}}. \qquad \left\{\begin{matrix}\text{Correlation coefficient}\\ \text{in terms of deviations}\\ \text{from means}\end{matrix}\right\} \qquad (32) $$

$$ r = \frac{\Sigma XY - NM_xM_y}{\sqrt{(\Sigma X^2 - NM_x^2)(\Sigma Y^2 - NM_y^2)}}. \qquad \left\{\begin{matrix}\text{Correlation coefficient}\\ \text{(based on raw scores)}\end{matrix}\right\} \qquad (33) $$

$$ r = \frac{\Sigma XY - T_xM_y}{\sqrt{(\Sigma X^2 - T_xM_x)(\Sigma Y^2 - T_yM_y)}}. \qquad \left\{\begin{matrix}\text{Correlation coefficient}\\ \text{equivalent to (33)}\end{matrix}\right\} \qquad (34) $$

$$ r = \frac{\Sigma f_{xy}d_xd_y - \dfrac{(\Sigma f_xd_x)\,(\Sigma f_yd_y)}{N}}{\sqrt{\left[\Sigma f_xd_x^2 - \dfrac{(\Sigma f_xd_x)^2}{N}\right]\left[\Sigma f_yd_y^2 - \dfrac{(\Sigma f_yd_y)^2}{N}\right]}} = \frac{a}{\sqrt{bc}}. \qquad (35) $$

{Correlation coefficient for distribution table}

$$ \bar{y} = \left(r\frac{\sigma_y}{\sigma_x}\right)x. \qquad \left\{\begin{matrix}\text{Regression line for}\\ \text{means of columns, re-}\\ \text{ferred to mean of table}\end{matrix}\right\} \qquad (36) $$

$$ S_y = \sigma_y\sqrt{1 - r^2}. \qquad \left\{\begin{matrix}\text{Standard error}\\ \text{of estimate}\end{matrix}\right\} \qquad (37) $$

$$ \bar{x} = \left(r\frac{\sigma_x}{\sigma_y}\right)y. \qquad \left\{\begin{matrix}\text{Regression line for}\\ \text{means of rows, referred}\\ \text{to mean of table}\end{matrix}\right\} \qquad (38) $$

$$ \bar{Y} = r\frac{\sigma_y}{\sigma_x}X - r\frac{\sigma_y}{\sigma_x}M_x + M_y, \qquad (39) $$

$$ \bar{X} = r\frac{\sigma_x}{\sigma_y}Y - r\frac{\sigma_x}{\sigma_y}M_y + M_x. \qquad \left\{\begin{matrix}\text{Regression}\\ \text{lines in score}\\ \text{form}\end{matrix}\right\} \qquad (40) $$

$$ \bar{Y} = \frac{ak}{bh}X - \frac{ak}{bh}M_x + M_y, \qquad (41) $$

$$ \bar{X} = \frac{ah}{ck}Y - \frac{ah}{ck}M_y + M_x. \qquad \left\{\begin{matrix}\text{Regression lines in}\\ \text{score form and sym-}\\ \text{bols on correlation}\\ \text{sheet}\end{matrix}\right\} \qquad (42) $$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{ak}{bh}, \tag{43}$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{ah}{ck}. \tag{44}$$

{Regression coefficients}

$$P.E.\,(\text{est. } Y) = .6745\,\sigma_y\sqrt{1-r^2}, \tag{45}$$

{Probable error of estimate in predicting Y from X}

$$P.E.\,(\text{est. } X) = .6745\,\sigma_x\sqrt{1-r^2}. \tag{46}$$

{Probable error of estimate in predicting X from Y}

$$I_p = 100\left(\frac{\sigma - \sigma\sqrt{1-r^2}}{\sigma}\right) = 100\left(1-\sqrt{1-r^2}\right). \tag{47}$$

{Improvement over chance in prediction by a single score}

$$r_{nn} = \frac{nr_{11}}{1+(n-1)r_{11}}. \tag{48}$$

{Spearman-Brown formula for predicting reliability of lengthened tests}

$$r_{cn} = \frac{nr_{cz}}{\sqrt{n+n(n-1)r_{zz}}}. \tag{51}$$

{Formula for predicting validity of lengthened tests}

$$S = R - \frac{1}{(n-1)}W = R - CW. \tag{52}$$

{Multiple-response scoring formula}

$$R_{12} = \frac{\Sigma_1}{\sigma_1}\frac{r_{12}}{\sqrt{1-r_{12}^2+r_{12}^2\left(\frac{\Sigma_1}{\sigma_1}\right)^2}}. \tag{53}$$

{Correlation after selection}

$$\eta_{yx} = \sqrt{1-\frac{\sigma_{ay}^2}{\sigma_y^2}}, \tag{55}$$

$$\eta_{xy} = \sqrt{1-\frac{\sigma_{ax}^2}{\sigma_x^2}}. \tag{56}$$

{Correlation ratios, original form}

$$\eta_{yx} = \frac{\sigma_{\bar{y}_x}}{\sigma_y}, \tag{60}$$

$$\eta_{xy} = \frac{\sigma_{\bar{x}_y}}{\sigma_x}. \tag{61}$$

{Correlation ratios as quotients of two standard deviations}

$$\eta_{yx} = \frac{\sqrt{\frac{\Sigma f_x(M_y-\bar{Y}_x)^2}{N}}}{\sigma_y}. \tag{62}$$

{Correlation ratio for means of columns}

$$\eta_{yx} = \sqrt{\frac{e}{c}}, \qquad (63)$$

{Correlation ratios for correlation blank}

$$\eta_{xy} = \sqrt{\frac{d}{b}}. \qquad (64)$$

$$\overline{X}_y = A_x + \left(\frac{\Sigma' f_{xy} d_x}{f_y}\right) h, \qquad (65)$$

$$\overline{Y}_x = A_y + \left(\frac{\Sigma' f_{xy} d_y}{f_x}\right) k. \qquad (66)$$

{Means of the arrays in a correlation table}

$$\eta^2 - r^2 < \frac{4.047}{\sqrt{N}} \sqrt{(\eta^2 - r^2)\{(1-\eta^2)^2 - (1-r^2)^2 + 1\}}. \qquad (67)$$

{Blakeman's test for linearity}

$$\sqrt{N}\sqrt{\eta^2 - r^2} < 4.047. \qquad (68)$$

{Blakeman's short test for linearity}

$$Y_s = \overline{Y}_s + (Y_t - \overline{Y}_t). \qquad (69)$$

{Corrective formula for eliminating age}

$$Y_s = \overline{Y}_s + (Y_t - \overline{Y}_t)\frac{\sigma_s}{\sigma_t}. \qquad (70)$$

{Corrective formula adjusting for age and heteroscedasticity}

$${}_nP_r = n(n-1)(n-2)\cdots(n-r+1). \qquad (71)$$

{Permutation of n things r at a time}

$${}_nC_r = \frac{n(n-1)(n-2)\cdots(n-r+1)}{1\cdot 2\cdot 3\cdots r} = \frac{{}_nP_r}{r!}. \qquad (72)$$

{Combination of n things r at a time}

$$(q+p)^n = {}_nC_0 q^n + {}_nC_1 q^{n-1}p + {}_nC_2 q^{n-2}p^2 + {}_nC_3 q^{n-3}p^3 + \cdots + {}_nC_n p^n. \qquad (77)$$

{Point binomial}

$$M = np. \qquad (78)$$

{Mean of the point binomial}

$$\sigma = \sqrt{npq}. \qquad (79)$$

{Standard deviation of the point binomial}

$$y = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}. \qquad (80)$$

{Normal curve with area = 1}

$$y = \frac{N}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} = y_0 e^{-\frac{x^2}{2\sigma^2}}. \qquad (82)$$

{Normal curve with area = N}

$$z = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \qquad \left\{\begin{array}{c}\text{Ordinate of the normal curve, with}\\ \text{unit area and standard deviation}\end{array}\right\} \qquad (83)$$

$$P.E. = .6744898\,\sigma. \qquad \left\{\begin{array}{c}\text{Relation between}\\ P.E. \text{ and } \sigma\end{array}\right\} \qquad (84)$$

$${}_1\bar{x}_2 = \frac{z_1 - z_2}{{}_1n_2}. \qquad \left\{\begin{array}{c}\text{Mean of a portion of a normal curve,}\\ \text{with unit area and standard deviation}\end{array}\right\} \qquad (85)$$

$$\frac{{}_1\bar{x}_2}{\sigma} = \frac{z_1 - z_2}{\frac{{}_1f_2}{N}}. \qquad \left\{\begin{array}{c}\text{Mean of a portion}\\ \text{of a normal curve,}\\ \text{with area} = N\end{array}\right\} \qquad (86)$$

$$P.E._M = .6745\,\frac{\sigma_x}{\sqrt{N}}. \qquad \left\{\begin{array}{c}\text{Probable error}\\ \text{of the mean}\end{array}\right\} \qquad (88)$$

$$P.E._{M_1 - M_2} = \sqrt{(P.E._{M_1})^2 + (P.E._{M_2})^2}. \qquad \left\{\begin{array}{c}\text{Probable error of the}\\ \text{difference between two}\\ \text{uncorrelated means}\end{array}\right\} \qquad (89)$$

$$P.E._{Md} = \frac{.84535\,\sigma_x}{\sqrt{N}} = 1.2533\,P.E._M. \qquad \left\{\begin{array}{c}\text{Probable error}\\ \text{of the median}\end{array}\right\} \qquad (90)$$

$$P.E._\sigma = \frac{.6745\,\sigma}{\sqrt{2N}} = \frac{.4769\,\sigma}{\sqrt{N}} = .7071\,P.E._M. \qquad \left\{\begin{array}{c}\text{Probable error of the}\\ \text{standard deviation}\end{array}\right\} \qquad (91)$$

$$P.E._V = \left\{\frac{.6745}{\sqrt{2N}}\right\}\left\{V\left[1 + 2\left(\frac{V}{100}\right)^2\right]^{\frac{1}{2}}\right\} = \chi_2\psi. \qquad \left\{\begin{array}{c}\text{Probable error}\\ \text{of } V \text{ with Pear-}\\ \text{son's Tables}\end{array}\right\} \qquad (93)$$

$$P.E._r = \frac{.6745\,(1 - r^2)}{\sqrt{N}}. \qquad \left\{\begin{array}{c}\text{Probable error of the}\\ \text{correlation coefficient}\end{array}\right\} \qquad (94)$$

$$P.E._\eta = \frac{.6745\,(1 - \eta^2)}{\sqrt{N}}. \qquad \left\{\begin{array}{c}\text{Probable error of the}\\ \text{correlation ratio}\end{array}\right\} \qquad (95)$$

$$P.E._{b_{xy}} = .6745\,\frac{\sigma_x}{\sigma_y}\,\frac{\sqrt{1 - r^2}}{\sqrt{N}}, \qquad (96a)$$

$$P.E._{b_{yx}} = .6745\,\frac{\sigma_y}{\sigma_x}\,\frac{\sqrt{1 - r^2}}{\sqrt{N}}. \qquad \left\{\begin{array}{c}\text{Probable errors}\\ \text{of regression co-}\\ \text{efficients}\end{array}\right\} \qquad (96b)$$

$$P.E._{b_{12.k}} = .6745\,\frac{\sigma_{1.2k}}{\sigma_{2.k}\sqrt{N}}. \quad \left\{\begin{array}{c}\text{Probable error of higher-order}\\ \text{regression coefficient}\end{array}\right\} \quad (97)$$

$$P.E._{\delta} = \frac{2(.6745)}{\sqrt{N}}\sqrt{(\eta^2 - r^2)\{(1-\eta^2)^2 - (1-r^2)^2 + 1\}}. \quad (98)$$

{Probable error of $\eta^2 - r^2$}

$$P.E._{A-B} = \sqrt{(P.E._A)^2 + (P.E._B)^2 - 2\,R_{AB}(P.E._A)(P.E._B)}. \quad (99)$$

{Probable error of difference with correlated measures}

$$P.E._{M_1 - M_2} = \sqrt{(P.E._{M_1})^2 + (P.E._{M_2})^2 - 2\,r_{12}\,P.E._{M_1}\,P.E._{M_2}}. \quad (100)$$

{Probable error of difference between means where correlated}

$$P.E._f = .6745\sqrt{f\left(1 - \frac{f}{N}\right)}. \quad \left\{\begin{array}{c}\text{Probable error of an}\\ \text{observed frequency}\end{array}\right\} \quad (101)$$

$$P.E._{f_p} = .6745\sqrt{\frac{f_p\,(100 - f_p)}{N}}. \quad \left\{\begin{array}{c}\text{Probable error of a per-}\\ \text{centage frequency}\end{array}\right\} \quad (102)$$

$$\chi^2 = \sum_{t=1}^{t=n}\left\{\frac{(f'_t - f_t)^2}{f_t}\right\}. \quad \left\{\begin{array}{c}\text{Chi-square}\\ \text{function}\end{array}\right\} \quad (103)$$

$$P.E._{np} = .6745\sqrt{npq}, \quad (105)$$

$$P.E._p = .6745\sqrt{\frac{pq}{n}}. \quad (106)$$

$\left\{\begin{array}{c}\text{Probable errors of the}\\ \text{mean and of the pro-}\\ \text{portion of successes}\end{array}\right\}$

$$P.E._{e_1}\ (\text{of individual } X_1) = .6745\,\sigma_{x_1}\sqrt{1 - r_{1I}}. \quad (111)$$

{Probable error of response for X_1}

$$P.E.\ (\text{of individual } z_1 - z_2) = .6745\sqrt{2 - r_{1I} - r_{2II}}. \quad (113)$$

$$r_{st} = \frac{r_{12}}{\sqrt{r_{1I}r_{2II}}}. \quad \left\{\begin{array}{c}\text{Spearman's correction}\\ \text{for attenuation}\end{array}\right\} \quad (115)$$

$$\frac{\sigma}{\Sigma} = \frac{\sqrt{1 - R_{1I}}}{\sqrt{1 - r_{1I}}}, \quad (116a)$$

$$r_{1I} = \frac{\sigma^2 - \Sigma^2(1 - R_{1I})}{\sigma^2}. \quad (116b)$$

$\left\{\begin{array}{c}\text{Kelley's formula for}\\ \text{adjusting reliability}\\ \text{coefficients}\end{array}\right\}$

$$r = \frac{\Sigma f_x d_x (\bar{Y}_x - M_y) h}{N \sigma_x \sigma_y}, \qquad \text{(118a)}$$

$$r = \frac{\Sigma f_y d_y (\bar{X}_y - M_x) k}{N \sigma_x \sigma_y}. \qquad \text{(118b)}$$

{Pearson's formulas for the correlation coefficient based on the means of the arrays}

$$r = \frac{\Sigma f_y d_y \left(\frac{\bar{x}_y}{\sigma_x}\right) k}{N \sigma_y}. \qquad \text{(119)}$$

{Correlation coefficient adapted for use with data on a normal scale}

$${}_c r_{xy} = \frac{\Sigma f_{xy} \frac{N}{f_x f_y} (z_s - z_{s+1})(z'_s - z'_{s+1})}{\left[\Sigma \frac{N}{f_x} (z_s - z_{s+1})^2\right]\left[\Sigma \frac{N}{f_y} (z'_s - z'_{s+1})^2\right]}. \qquad \text{(120)}$$

{Pearson's corrective formula for broad grouping assuming normal distributions of the variates}

$$\eta_{xy} = \frac{\sqrt{\frac{\Sigma f_y \bar{x}_y^2}{N}}}{\sigma_x} = \sqrt{\frac{\Sigma f_y \left(\frac{\bar{x}_y}{\sigma_x}\right)^2}{N}}. \qquad \text{(121)}$$

{Correlation ratio adapted for use with data on a normal scale}

$${}_c \eta_{yx} = \frac{\eta_{yx}}{r_{xc}}. \qquad \text{(122)}$$

{Correlation ratio corrected for broad categories}

$$r_{xc} = \sqrt{\Sigma \frac{N}{f_x} (z_s - z_{s+1})^2}. \qquad \text{(123)}$$

{Correlation of a variable with its class value}

$$r_{\text{bis.}} = \frac{\bar{Y}_2 - \bar{Y}_1}{\sigma_y} \left(\frac{pq}{z}\right). \qquad \text{(124)}$$

{Biserial r}

$$P.E._{(\text{bis. } r)} = \frac{.6745 \left(\sqrt{\frac{pq}{z^2}} - r^2\right)}{\sqrt{N}}. \qquad \text{(125)}$$

{Probable error of biserial r}

$$C = \sqrt{\frac{S' - N}{N + S' - N}} = \sqrt{\frac{S' - N}{S'}}, \qquad \text{(128a)}$$

{First computation form for contingency}

$$C = \sqrt{\frac{S - 1}{S}}. \qquad \text{(128b)}$$

{Second computation form for contingency}

$$_cC = \frac{C}{r_{xc}r_{yc}}. \qquad \left\{\begin{array}{c}\text{Correction to the con-}\\ \text{tingency coefficient for}\\ \text{broad grouping}\end{array}\right\} \quad (129)$$

$$P.E._c = \frac{.6745}{\sqrt{N}}\left[\frac{\frac{\psi^3}{\phi^2}+1-\phi^2}{(1+\phi^2)^3}\right]^{\frac{1}{2}}. \qquad \left\{\begin{array}{c}\text{Probable error}\\ \text{of contingency}\\ \text{coefficient}\end{array}\right\} \quad (130)$$

$$\rho = 1 - \frac{6\,\Sigma(v_x - v_y)^2}{N(N^2-1)}. \qquad \left\{\begin{array}{c}\text{Spearman's formula}\\ \text{based on rank dif-}\\ \text{ferences}\end{array}\right\} \quad (131)$$

$$P.E._\rho = \frac{.7063(1-\rho^2)}{\sqrt{N}}. \qquad \left\{\begin{array}{c}\text{Probable error of}\\ \text{Spearman's rank}\\ \text{coefficient}\end{array}\right\} \quad (133)$$

$$r_{12.3} = \frac{r_{12} - r_{13}\cdot r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}. \qquad \left\{\begin{array}{c}\text{Partial-correlation}\\ \text{coefficient for three}\\ \text{variables}\end{array}\right\} \quad (134)$$

$$r_{12.34\cdots n} = \frac{r_{12.34\cdots(n-1)} - r_{1\,n.34\cdots(n-1)}r_{2\,n.34\cdots(n-1)}}{\sqrt{[1-r^2_{1\,n.34\cdots(n-1)}][1-r^2_{2\,n.34\cdots(n-1)}]}}. \quad (135)$$

{Partial-correlation coefficient of the order $(n-2)$}

$$\bar{X}_1 = b_{12.34\cdots n}X_2 + b_{13.24\cdots n}X_3 + \cdots + b_{1\,n.23\cdots(n-1)}X_n + C. \quad (139)$$

{Regression equation for estimating X_1 from the remaining $(n-1)$ variables}

$$b_{12.34\cdots n} = r_{12.34\cdots n}\frac{\sigma_{1.34\cdots n}}{\sigma_{2.34\cdots n}}. \qquad \left\{\begin{array}{c}\text{Regression coefficient}\\ \text{of the order } (n-2)\end{array}\right\} \quad (140)$$

$$\sigma_{1.23\cdots n} = \sigma_1\sqrt{(1-r_{12}^2)(1-r_{13.2}^2)\cdots(1-r^2_{1\,n.23\cdots(n-1)})}. \quad (141)$$

{Standard deviation of the order $(n-1)$}

$$P.E._{est} = .6745\,\sigma_{1.23\cdots n}. \quad (142)$$

{Probable error of estimate}

$$C = M_1 - b_{12.34\cdots n}M_2 - b_{13.24\cdots n}M_3 - \cdots - b_{1\,n.23\cdots(n-1)}M_n. \quad (146)$$

{Constant term in regression equation}

$$\sigma_{1.23} = \sigma_1\sqrt{\frac{1-r_{12}^2-r_{13}^2-r_{23}^2+2\,r_{12}r_{13}r_{23}}{1-r_{23}^2}} = \frac{\sigma_1\sqrt{S_{123}}}{\sqrt{1-r_{23}^2}}. \quad (147)$$

{Standard error of second-order in terms of zero-order coefficients}

$$r_{12.34}=\frac{r_{12}(1-r_{34}^2)-r_{13}r_{23}-r_{14}r_{24}+r_{34}(r_{13}r_{24}+r_{14}r_{23})}{\sqrt{(1-r_{13}^2-r_{14}^2-r_{34}^2+2\,r_{13}r_{14}r_{34})(1-r_{23}^2-r_{24}^2-r_{34}^2+2\,r_{23}r_{24}r_{34})}}$$
$$=\frac{S_{12.34}}{\sqrt{S_{134}S_{234}}}, \qquad (151a)$$

$$r_{13.24}=\frac{r_{13}(1-r_{24}^2)-r_{12}r_{23}-r_{14}r_{34}+r_{24}(r_{12}r_{34}+r_{14}r_{23})}{\sqrt{(1-r_{12}^2-r_{14}^2-r_{24}^2+2\,r_{12}r_{14}r_{24})(1-r_{23}^2-r_{24}^2-r_{34}^2+2\,r_{23}r_{24}r_{34})}}$$
$$=\frac{S_{13.24}}{\sqrt{S_{124}S_{234}}}, \qquad (151b)$$

$$r_{14.23}=\frac{r_{14}(1-r_{23}^2)-r_{12}r_{24}-r_{13}r_{34}+r_{23}(r_{12}r_{34}+r_{13}r_{24})}{\sqrt{(1-r_{12}^2-r_{13}^2-r_{23}^2+2\,r_{12}r_{13}r_{23})(1-r_{23}^2-r_{24}^2-r_{34}^2+2\,r_{23}r_{24}r_{34})}}$$
$$=\frac{S_{14.23}}{\sqrt{S_{123}S_{234}}}. \qquad (151c)$$

{Second-order correlation coefficients in terms of zero-order coefficients}

$$b_{12.34\cdots n}$$
$$=\frac{r_{12.34\cdots(n-1)}-r_{1\,n.34\cdots(n-1)}r_{2\,n.34\cdots(n-1)}}{1-r_{2\,n.34\cdots(n-1)}^2}\frac{\sigma_{1.34\cdots(n-1)}}{\sigma_{2.34\cdots(n-1)}}. \qquad (152)$$

{Reduction formula for regression coefficient}

$$b_{12.34}=\frac{\sigma_1}{\sigma_2}\left[\frac{r_{12}(1-r_{34}^2)-r_{13}r_{23}-r_{14}r_{24}+r_{34}(r_{13}r_{24}+r_{14}r_{23})}{1-r_{23}^2-r_{24}^2-r_{34}^2+2\,r_{23}r_{24}r_{34}}\right]$$
$$=\frac{\sigma_1}{\sigma_2}\frac{S_{12.34}}{S_{234}}, \qquad (153a)$$

$$b_{13.24}=\frac{\sigma_1}{\sigma_3}\left[\frac{r_{13}(1-r_{24}^2)-r_{12}r_{23}-r_{14}r_{34}+r_{24}(r_{12}r_{34}+r_{14}r_{23})}{1-r_{23}^2-r_{24}^2-r_{34}^2+2\,r_{23}r_{24}r_{34}}\right]$$
$$=\frac{\sigma_1}{\sigma_3}\frac{S_{13.24}}{S_{234}}, \qquad (153b)$$

$$b_{14.23}=\frac{\sigma_1}{\sigma_4}\left[\frac{r_{14}(1-r_{23}^2)-r_{12}r_{24}-r_{13}r_{34}+r_{23}(r_{12}r_{34}+r_{13}r_{24})}{1-r_{23}^2-r_{24}^2-r_{34}^2+2\,r_{23}r_{24}r_{34}}\right]$$
$$=\frac{\sigma_1}{\sigma_4}\frac{S_{14.23}}{S_{234}}. \qquad (153c)$$

{Second-order regression coefficients in terms of zero-order coefficients}

$$\sigma_{1.234}=\sigma_1\sqrt{\frac{S_{123}S_{234}-S_{14.23}^2}{(1-r_{23}^2)S_{234}}}. \qquad \left\{\begin{array}{l}\text{Standard deviation of}\\ \text{third-order in terms of}\\ \text{zero-order coefficients}\end{array}\right\} \qquad (154)$$

$$R_{1(23\cdots n)} = r_{X_1\bar{X}_1} = \sqrt{1 - \frac{\sigma^2_{1.23\cdots n}}{\sigma_1^2}}. \quad \left\{\begin{matrix}\text{Multiple-correlation}\\ \text{coefficient}\end{matrix}\right\} \quad (155)$$

$$1 - R^2_{1(23\cdots n)} = (1 - r^2_{12})(1 - r^2_{13.2})(1 - r^2_{14.23})\cdots(1 - r^2_{1n.23\cdots(n-1)}). \quad (156)$$

{Computation form for R}

$$R_{1(23\cdots n)} = r_{1x}\sqrt{\frac{n-1}{1+(n-2)r_{xx}}}. \quad \left\{\begin{matrix}\text{Multiple-correlation}\\ \text{coefficient for equal}\\ \text{coefficients}\end{matrix}\right\} \quad (157)$$

$$\Delta = \begin{vmatrix} r_{11} & r_{21} & r_{31} & \cdots & r_{n1} \\ r_{12} & r_{22} & r_{32} & \cdots & r_{n2} \\ r_{13} & r_{23} & r_{33} & \cdots & r_{n3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{1n} & r_{2n} & r_{3n} & \cdots & r_{nn} \end{vmatrix}. \quad \left\{\begin{matrix}\text{Determinant}\\ \text{of zero-order}\\ \text{coefficients}\end{matrix}\right\} \quad (160)$$

$$\Delta = \begin{vmatrix} r_{11} & r_{21} & r_{31} \\ r_{12} & r_{22} & r_{32} \\ r_{13} & r_{23} & r_{33} \end{vmatrix}. \quad \left\{\begin{matrix}\text{Determinant for}\\ \text{three variables}\end{matrix}\right\} \quad (161)$$

$$\Delta = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\, r_{12}r_{13}r_{23}. \quad \left\{\begin{matrix}\text{Expanded}\\ \text{value of (161)}\end{matrix}\right\} \quad (162)$$

$$r_{12.34\cdots n} = \frac{-A_{12}}{\sqrt{\Delta_{11}\Delta_{22}}}. \quad (163)$$

$$r_{1\,k.34\cdots k\cdots n} = \frac{-A_{1k}}{\sqrt{\Delta_{11}\Delta_{kk}}}. \quad (164)$$

$$R_{1(23\cdots n)} = \sqrt{1 - \frac{\Delta}{\Delta_{11}}}. \quad (165)$$

$$\sigma_{1.23\cdots n} = \sigma_1\sqrt{1 - R^2} = \sigma_1\sqrt{\frac{\Delta}{\Delta_{11}}}. \quad (166)$$

$$b_{12.34\cdots n} = \frac{-A_{12}}{\Delta_{11}}\frac{\sigma_1}{\sigma_2}. \quad (167)$$

$$b_{1\,k.23\cdots k\cdots n} = \frac{-A_{1k}}{\Delta_{11}}\frac{\sigma_1}{\sigma_k}. \quad (168)$$

$$Y = \frac{X}{a + bX} + c. \qquad \{\text{Hyperbola}\} \qquad (171)$$

$$Y = a + bX + c \log X. \quad \{\text{Logarithmic growth function}\} \quad (172)$$

$$Y = C_0 + C_1X + C_2X^2 + C_3X^3 + \cdots + C_nX^n. \quad \{n\text{th-order parabola}\} \quad (173)$$

$$\Sigma Y = C_0\Sigma(1) + C_1\Sigma(X) + C_2\Sigma(X^2) + \cdots + C_n\Sigma(X^n), \qquad (175a)$$

$$\Sigma XY = C_0\Sigma(X) + C_1\Sigma(X^2) + C_2\Sigma(X^3) + \cdots + C_n\Sigma(X^{n+1}), \qquad (175b)$$

$$\Sigma X^2Y = C_0\Sigma(X^2) + C_1\Sigma(X^3) + C_2\Sigma(X^4) + \cdots + C_n\Sigma(X^{n+2}), \qquad (175c)$$

. . . . . .

$$\Sigma X^nY = C_0\Sigma(X^n) + C_1\Sigma(X^{n+1}) + C_2\Sigma(X^{n+2}) + \cdots + C_n\Sigma(X^{2n}). \qquad (175d)$$

{Normal equations, unweighted ordinates}

$$\Sigma f_xY = C_0\Sigma(f_x) + C_1\Sigma(f_xX) + C_2\Sigma(f_xX^2) + \cdots + C_n\Sigma(f_xX^n), \qquad (176a)$$

$$\Sigma f_xXY = C_0\Sigma(f_xX) + C_1\Sigma(f_xX^2) + C_2\Sigma(f_xX^3) + \cdots + C_n\Sigma(f_xX^{n+1}), \qquad (176b)$$

. . . . . .

$$\Sigma f_xX^nY = C_0\Sigma(f_xX^n) + C_1\Sigma(f_xX^{n+1}) + C_2\Sigma(f_xX^{n+2}) + \cdots + C_n\Sigma(f_xX^{2n}). \qquad (176c)$$

{Normal equations, weighted ordinates}

$$\Sigma(Y) = a\Sigma(1) + b\Sigma(X) + c\Sigma(\log X), \qquad (183a)$$

$$\Sigma(XY) = a\Sigma(X) + b\Sigma(X^2) + c\Sigma(X \log X), \qquad (183b)$$

$$\Sigma(Y \log X) = a\Sigma(\log X) + b\Sigma(X \log X) + c\Sigma(\log X)^2. \qquad (183c)$$

{Normal equations for the logarithmic growth curve}

$$p\text{th moment about the origin} = \Sigma fX^p. \qquad (188)$$

$$\bar{v}_p = \frac{\Sigma fX^p}{N}. \qquad \left\{\begin{matrix}\text{Moment coefficient}\\ \text{about the origin}\end{matrix}\right\} \qquad (189)$$

$$v_p = \frac{\Sigma fx^p}{N} = \frac{\Sigma f(X - M)^p}{N}. \qquad \left\{\begin{matrix}\text{Formula for mo-}\\ \text{ment coefficients}\\ \text{about the mean}\end{matrix}\right\} \qquad (190)$$

$$\nu_0 = \frac{\Sigma f x^0}{N} = 1, \tag{191a}$$

$$\nu_1 = \frac{\Sigma f x}{N} = 0, \tag{191b}$$

$$\nu_2 = \frac{\Sigma f x^2}{N} = \sigma_x{}^2, \text{ etc.} \tag{191c}$$

{Moment coefficients about the mean}

$$\nu_1 = \bar{\nu}_1 - \bar{\nu}_1 = 0, \tag{193a}$$

$$\nu_2 = \bar{\nu}_2 - \bar{\nu}_1{}^2, \tag{193b}$$

$$\nu_3 = \bar{\nu}_3 - 3\,\bar{\nu}_1\bar{\nu}_2 + 2\,\bar{\nu}_1{}^3, \tag{193c}$$

$$\nu_4 = \bar{\nu}_4 - 4\,\bar{\nu}_1\bar{\nu}_3 + 6\,\bar{\nu}_1{}^2\bar{\nu}_2 - 3\,\bar{\nu}_1{}^4. \tag{193d}$$

{Moment coefficients about the mean in terms of those about the origin}

$$\nu_0 = 1, \tag{195a}$$

$$\nu_1 = 0, \tag{195b}$$

$$\nu_2 = \bar{\nu}_2 - \bar{\nu}_1{}^2, \tag{195c}$$

$$\nu_3 = \bar{\nu}_3 - 3\,\bar{\nu}_1\nu_2 - \bar{\nu}_1{}^3, \tag{195d}$$

$$\nu_4 = \bar{\nu}_4 - 4\,\bar{\nu}_1\nu_3 - 6\,\bar{\nu}_1{}^2\nu_2 - \bar{\nu}_1{}^4. \tag{195e}$$

{Moment coefficients about the mean in terms of moments about the origin and mean}

$$\mu_1 = \nu_1, \tag{196a}$$

$$\mu_2 = \nu_2 - \tfrac{1}{12} = \nu_2 - .083333, \tag{196b}$$

$$\mu_3 = \nu_3, \tag{196c}$$

$$\mu_4 = \nu_4 - \tfrac{1}{2}\nu_2 + \tfrac{7}{240} = \nu_4 - .5\,\nu_2 + .02916667. \tag{196d}$$

{Moment coefficients about the mean adjusted for grouping (Sheppard's corrections)}

$$\beta_1 = \frac{\mu_3{}^2}{\mu_2{}^3}, \qquad \kappa_1 = 2\,\beta_2 - 3\,\beta_1 - 6,$$

$$\beta_2 = \frac{\mu_4}{\mu_2{}^2}, \qquad \kappa_2 = \frac{\beta_1(\beta_2+3)^2}{4\,(4\,\beta_2 - 3\,\beta_1)(2\,\beta_2 - 3\,\beta_1 - 6)}. \tag{197}$$

{Pearson's constants for curve-fitting}

APPENDIX B

LIST OF IMPORTANT SYMBOLS

In the following list the symbols are given in the order in which they first appear in the formulas of Appendix A.

1. M denotes the arithmetic mean.
2. Σ denotes the sum of the items of the sort indicated.
3. X denotes a raw score taken as a deviation from zero.
4. N denotes the size of the sample or the number of cases used.
5. A denotes an assumed mean or arbitrary origin.
6. f denotes the frequency in a class interval.
7. d denotes a score as a deviation from an assumed mean and is expressed in units of class intervals.
8. h denotes the width of the class interval.
9. Md denotes the median.
10. $l.l.$ denotes the lower limit of the interval containing the median in formula (8 a).
11. $u.l.$ denotes the upper limit of the interval containing the median in formula (8 b).
12. f_{up} denotes the total frequency up to the interval containing the median.
13. f_{do} denotes the total frequency down to the interval containing the median.
14. f_{md} denotes the frequency in the interval containing the median.
15. $G.M.$ denotes the geometric mean.
16. $X_1X_2X_3 \cdots X_N$ denotes the product of the N values of X.
17. H denotes the harmonic mean.
18. $M.D.$ denotes the mean deviation of scores from the arithmetic mean.
19. $|x|$ denotes the absolute value of x, where $x = X - M$.
20. A_m denotes the mid-point of the interval in which M lies.
21. N_a denotes the number of cases above M.
22. N_b denotes the number of cases below M.
23. $S.D.$ denotes the standard deviation.

24. X' denotes a deviation of the score from an assumed mean, that is, $X' = X - A$.

25. M' denotes the arithmetic mean of the X' scores.

26. Q denotes the quartile deviation.

27. Q_1 denotes the first quartile, which is the value below which one quarter of the cases lie.

28. Q_3 denotes the third quartile, which is the value below which three quarters of the cases lie.

29. $u.l.$ denotes the upper limit of the interval containing Q_3 in formula (20 a).

30. $l.l.$ denotes the lower limit of the interval containing Q_1 in formula (20 b).

31. f_{do} denotes the total frequency down to the interval containing Q_3 in formula (20 a).

32. f_{up} denotes the total frequency up to the interval containing Q_1 in formula (20 b).

33. f_3 denotes the frequency in the interval containing Q_3.

34. f_1 denotes the frequency in the interval containing Q_1.

35. V denotes the coefficient of variation.

36. σ denotes the standard deviation.

37. S_k denotes a measure of skewness.

38. M_o denotes the mode.

39. P_p denotes a percentile value.

40. p denotes the percentage of the cases smaller than P_p in formulas (27 a) and (27 b).

41. f_P denotes the frequency in the interval where P_p lies.

42. R_x denotes the percentile rank of a score X in formulas (28) and (29).

43. R_l denotes the percentile rank of the lower limit of the interval containing X.

44. R_u denotes the percentile rank of the upper limit of the interval containing X.

45. f_x denotes the frequency in the interval containing X in formula (29).

46. ${}_cR_x$ denotes the percentile rank of the middle of the interval containing X.

47. r denotes the product-moment coefficient of correlation.

48. x and y denote deviations from the respective means for X and Y, that is, $x = X - M_x$ and $y = Y - M_y$.

49. σ_x and σ_y denote the standard deviations for X and Y, respectively.

50. T_x and T_y denote the total of the X scores and the Y scores, respectively, that is, $T_x = \Sigma X$ and $T_y = \Sigma Y$.

51. f_x denotes the frequency of a column of type x in formula (35).

52. f_y denotes the frequency of a row of type y.

53. f_{xy} denotes the frequency of a cell common to a column and a row.

54. d_x and d_y denote the deviations in class intervals from the assumed means for the two variables.

55. a, b, and c denote the three parts of formula (35) and are defined by that equation.

56. S_y denotes the standard error of estimate in predicting Y from X by a regression equation.

57. $\bar{y}$ and $\bar{x}$ denote points on the regression lines (36) and (38), respectively.

58. $\overline{Y}$ and $\overline{X}$ denote points on the regression lines (39) and (40), respectively. It should be noted that $\bar{y} = \overline{Y} - M_y$ and $\bar{x} = \overline{X} - M_x$.

59. h and k denote the widths of the class intervals for X and Y, respectively.

60. b_{yx} denotes the regression coefficient for y on x.

61. b_{xy} denotes the regression coefficient for x on y.

62. $P.E.$ denotes the probable error.

63. I_p denotes the improvement over chance in prediction from a regression equation by a single score.

64. r_{nn} denotes the predicted reliability coefficient of a test n times its original length.

65. r_{1I} denotes the reliability coefficient or correlation between two parallel forms of a test.

66. r_{cn} denotes the correlation between a criterion and a test n times its original length.

67. r_{cz} denotes the average correlation between a criterion and each of several tests z_1, z_2, $z_3 \cdots z_n$ in formula (51).

68. r_{zz} denotes the average intercorrelation of the tests z_1, z_2, $z_3 \cdots z_n$ in formula (51).

69. S denotes the score corrected for guessing by formula (52).

70. R denotes the number of right responses in formula (52).

71. W denotes the number of wrong responses in formula (52).

72. C denotes a constant in formula (52).

73. n denotes the number of choices in answering a multiple-response test in formula (52).

74. R_{12} denotes the correlation after selection in formula (53).

75. Σ_1 denotes the standard deviation after selection in formula (53).

76. η_{yx} denotes the correlation ratio based on the means of the columns.

77. η_{xy} denotes the correlation ratio based on the means of the rows.

78. σ_{ay} denotes the standard deviation of $y - \bar{y}_x$, where $\bar{y}_x$ denotes the mean of a column.

79. $\sigma_{\bar{y}_x}$ denotes the standard deviation of $\bar{y}_x$.

80. $\bar{Y}_x$ denotes the mean of a column. It should be noted that $\bar{y}_x = \bar{Y} - M_y$.

81. e is defined by $\Sigma f_x(M_y - \bar{Y}_x)^2/k^2$.

82. d is defined by $\Sigma f_y(M_x - \bar{X}_y)^2/h^2$.

83. Σ' denotes summations over an array, for example, over a row or column in the correlation table.

84. A_x and A_y denote the assumed means for X and for Y, respectively.

85. Y_s denotes a variable corrected for age by formulas (69) and (70).

86. $\bar{Y}_s$ denotes the mean at age A_s in formulas (69) and (70).

87. σ_s denotes the standard deviation of the array at age A_s in formula (70).

88. ${}_nP_r$ denotes the permutation of n things r at a time.

89. ${}_nC_r$ denotes the combination of n things r at a time.

90. q denotes the probability for the failure of an event.

91. p denotes the probability for the success of an event.

92. n denotes the number of independent events for formula (77).

93. π denotes the value obtained by dividing the circumference of a circle by its radius.

94. e denotes the base of the Napierian system of logarithms as used in formula (80).

95. y_0 denotes the ordinate at $x = 0$ for a normal curve.

96. z denotes the ordinate of a normal curve with unit area and unit standard deviation.

97. ${}_1\bar{x}_2$ denotes the mean of the portion of a normal curve lying between the ordinates z_1 and z_2.

98. ${}_1n_2$ denotes the fractional part of the area of a normal curve lying between the ordinates z_1 and z_2.

99. f_p denotes a percentage frequency.

100. χ^2 denotes the chi-square function given by formula (103).

101. f'_t and f_t denote observed and theoretical frequencies, respectively, in formula (103).

102. e_1 denotes response error in formula (111).

103. z_1 and z_2 denote standard scores defined by $\frac{x_1}{\sigma_1}$ and $\frac{x_2}{\sigma_2}$, respectively, in formula (113).

104. r_{st} denotes the correlation between "true" scores s and t which are freed from the influence of response error.

105. ${}_c r_{xy}$ denotes the correlation coefficient corrected for broad grouping.

106. z_s and z'_s denote ordinates on the two scales of a normal correlation surface as used in formula (120).

107. ${}_c\eta_{yx}$ denotes the correlation ratio corrected for broad grouping.

108. r_{xc} denotes the correlation of a variable with its class value.

109. q and p denote the parts of the unit normal curve to the left and right of the ordinate z in formula (124).

110. $r_{\text{bis.}}$ denotes biserial r.

111. C denotes the contingency coefficient in formulas (128a) and (128b).

112. ${}_c C$ denotes the contingency coefficient corrected for broad grouping.

113. S' is defined by $\Sigma\left\{\frac{f_{xy}^2}{\frac{f_x f_y}{N}}\right\}$ in formula (128a).

114. S is defined by $\Sigma\left\{\frac{f_{xy}^2}{f_x f_y}\right\}$ in formula (128b).

115. ϕ^2 is defined by $\frac{1}{N}\Sigma\left[\frac{\left(f_{xy}-\frac{f_x f_y}{N}\right)^2}{\frac{f_x f_y}{N}}\right] = S - 1$ in formula (130).

116. ψ^3 is defined by $\frac{1}{N}\Sigma\left[\frac{\left(f_{xy}-\frac{f_x f_y}{N}\right)^3}{\left(\frac{f_x f_y}{N}\right)^2}\right]$ in formula (130).

117. ρ denotes Spearman's rank difference correlation coefficient.

118. v_x and v_y denote the ranks for the X and the Y series, respectively, in formula (131).

119. $r_{12.34\ldots n}$ denotes a partial-correlation coefficient of the order $(n-2)$.

120. $r^2_{1n.34\ldots(n-1)}$ denotes the square of the designated partial-correlation coefficient.

121. $b_{12.34\ldots n}$ denotes a regression coefficient of the order $(n-2)$.

122. C denotes the constant term in a regression equation and is defined by formula (146).

123. $\sigma_{1.23\ldots n}$ denotes the standard deviation of the order $(n-1)$.

124. S_{123}, $S_{12.34}$, etc. denote sums involving zero-order correlation coefficients and are defined by formulas (147) and (151).

125. $R_{1(23\ldots n)}$ denotes the multiple-correlation coefficient.

126. Δ denotes a determinant.

127. Δ_{ij} denotes a minor of a determinant obtained by deleting the coefficients in a row and column common to r_{ij}.

128. A_{ij} denotes a cofactor and is equal to Δ_{ij} with the sign that would be attached in expanding the determinant.

129. $\bar{\nu}_p$ denotes a moment coefficient about the origin.

130. ν_p denotes a moment coefficient about the mean.

131. μ_1, μ_2, μ_3, and μ_4 denote adjusted moment coefficients about the mean.

132. β_1 and β_2 denote functions of the adjusted moment coefficients and are used in curve-fitting.

133. κ_1 and κ_2 denote functions of β_1 and β_2 and are used in curve-fitting.

APPENDIX C

SELECTED BOOKS FOR SUPPLEMENTARY READING

A. Texts on Educational Statistics:

1. *Statistical Methods Applied to Educational Problems*, by Harold O. Rugg. Houghton Mifflin Company, 1917.
   A very readable book on elementary methods.
2. *Statistics in Education and Psychology*, by Henry E. Garrett. Longmans, Green & Co., 1926.
   A good discussion of reliability and partial correlation.
3. *Fundamentals of Statistics*, by L. L. Thurstone. The Macmillan Company, 1925.
   A clear presentation of elementary methods.
4. *Statistical Method in Educational Measurement*, by Arthur S. Otis. World Book Company, 1925.
   Contains a full treatment of percentile curves.
5. *Statistical Method*, by T. L. Kelley. The Macmillan Company, 1923.
   An advanced book including many important formulas.
6. *Essentials of Mental Measurement*, by W. Brown and G. Thomson. Cambridge University Press, London, 1921.
   Discusses psychophysical methods and the Spearman two-factor theory.
7. *Graphic Methods in Education*, by J. H. Williams. Houghton Mifflin Company, 1924.
   Shows how to prepare charts and diagrams.

B. General Texts:

1. *Introduction to the Theory of Statistics*, by G. Yule. Charles Griffin, London, 1926.
   The best general text, but somewhat difficult for beginners.
2. *First Course in Statistics*, by D. C. Jones. G. Bell, London.
   A clearly written text. Contains a good discussion of frequency curve-fitting.

3. *Handbook of Mathematical Statistics*, by H. L. Rietz and others. Houghton Mifflin Company, 1924.
   A useful reference book.
4. *Mathematical Analysis of Statistics*, by C. H. Forsyth. John Wiley & Sons, 1924.
   Clear treatment of interpolation. Suitable for students with mathematical training.
5. *Mathematical Theory of Probabilities*, by Arne Fisher. The Macmillan Company, 1922.
   A careful development of the theory of probability and applications to statistical problems. For advanced students.
6. *Frequency Curves and Correlation*, by W. P. Elderton. C. and E. Layton, London, 1927.
   A good exposition of Pearson's System of frequency curve-fitting.
7. *Calculus of Observations*, by E. T. Whittaker and G. Robinson. D. Van Nostrand Company, 1924.
   An excellent text for the advanced mathematical student.
8. *Mathematical Statistics*, by Henry Lewis Rietz. The Open Court Publishing Company, Chicago, 1927.
   A concise, clear, and excellent monograph. Especially recommended for students who have had calculus.

C. Texts in Other Fields:

1. *Medical Biometry and Statistics*, by Raymond Pearl. W. B. Saunders Company, 1923.
   A clearly written text for students of medicine and public health.
2. *Statistical Methods*, by Frederick C. Mills. Henry Holt and Company, 1924.
   One of the best books in the field of economics.
3. *Elements of Statistics*, by A. L. Bowley. P. S. King, London, 1920.
   An advanced book on economic statistics, by the most authoritative writer.

D. Aids in Calculation:

1. *Tables for Statisticians and Biometricians*, edited by Karl Pearson. Cambridge University Press, London, 1924.
   New edition forthcoming.
   The best tables for advanced work.

2. *Tables of* $\sqrt{1 - r^2}$ *and* $1 - r^2$, by J. R. Miner. The Johns Hopkins Press, 1922.
   Every student with access to a calculation machine should have these tables.
3. *Barlow's Tables of Squares*, etc. (1–10,000). E. and F. Spar, London. (May be obtained at The University of Chicago Bookstore.)
   The classical handbook.
4. *Tables of Applied Mathematics in Statistics*, by J. W. Glover. George Wahr, Ann Arbor, Michigan, 1924.
   A valuable aid for the actuary and advanced student.
5. *Statistical Tables for Students in Education and Psychology*, by Karl J. Holzinger. The University of Chicago Press, 1925.
   Adapted for classroom use.
6. *Probable Errors of the Correlation Coefficient*, by Karl J. Holzinger. Cambridge University Press, London, 1925.
   Four-place values with proportional parts.
7. *Chambers's Mathematical Tables.* W. R. Chambers, London, 1921.
   Contains seven-place logarithm tables.
8. *Five-Place Logarithmic and Trigonometric Tables*, by James M. Taylor. Ginn and Company, 1905.
   A clearly printed and convenient set of tables.

INDEX

Age, corrective formula for eliminating, 185 f.
Analysis of classified data, 7
Area of normal curve, 209 f.
Arithmetical mean, 48; calculation of, 79 ff.; properties of, 83 f.; reliability of, 85
Arithmetical progression, 47
Attenuation, Spearman's correction for, 253
Averages, 78 ff.; method of, in curve-fitting, 320 f., 325 ff. *See also* Arithmetical mean, Geometrical mean, Harmonic mean, Median, Mode
Ayres, Leonard P., 10, 26

Bar diagrams, 38 f.
Binomial distribution, 190 ff.
Binomial law, experimental verification of, 199 f.
Biserial *r*, 271 ff.
Blakeman's test for linearity, 183, 267
Burgess, William R., 11 f., 299
Burt, Cyril, 305 ff.

Calculation, of statistical constants, 7; errors in, 65 ff.
Card, data, 20
Central tendency, variations in, 79
Characters, in statistical series, 12 f.; ordered and unordered, 13; continuous and discontinuous, 14; classes of, 22; static and dynamic, 75; methods of correlation for two, 256 ff.
Chi-Square Test, Pearson's, 245 ff.
Class limits, 23 f.
Class values, 24, 80; percentile rank of, 138 f.
Classification of data, 9 ff.
Classifier, 25 ff.
Coefficient, of variation, 116 ff.; product-moment correlation, 143 ff.; regression, 159, 161; reliability, 168 ff.; validity, 168; probable error of, of variation, 238; probable error of correlation, 238; of contingency, 273 ff.
Cofactor in a determinant, 312
Collection, units of, 11 f.
Column diagram, 36 f.
Combinations, 191
Comparable measurements, 118 ff.
Compensating errors, 66
Constants, statistical, 7
Contingency, coefficient of, 273 ff.
Coördinates, 40 ff.
Correlation, linear, 141 ff.; Spearman's theorem on, 168; Spearman-Brown prophecy formula, 169 f.; effect of selection upon, 172; non-linear, 177 ff.; methods of, for two characters, 256 ff.; partial, 283 ff.; multiple, 307 ff.
Correlation coefficient, product-moment, 143 ff.; computation of, 146 ff.; interpretation of, 163 ff.; probable error of, 238
Correlation ratio, 177 ff.; probable error of, 239; for qualitative and unordered series, 266 f.
Courses in experimental and statistical method, 2
Crude mode, 90 f.
Cumulative frequency curve, 129 ff.
Cumulative frequency distribution, for Otis Test, 28
Curve-fitting, elements of, 317 ff.
Curves, 44; normal probability, 44 f., 204 ff.; types of, 318 f.; fitting normal, by method of moments, 214 ff., 342 ff.; criteria and constants for normal, 343

Data, in statistical investigation, 3 f.; collection and analysis of, 6 f.; collection and classification of, 9 ff.; primary and secondary, 9 f.; methods of collecting, 14 ff.; arrangement of, 19 ff.; range of, 22; tabular and graphical presentation of, 31 ff.; cal-

culation of mean for ungrouped, 79; fitting normal curve to frequency distribution of, 214 ff.; representing, on a normal scale, 221 ff.
Data card, 20
Determinants, solution by, 312 ff.
Deviates of normal curve, 209 f.
Deviation, mean, 102 ff.; standard, 108 ff.; quartile, 110 ff.
Diagrams, presentation of results in, 7 f.; purpose of, 31 f.; column and bar, 36 ff.; scatter, 141 f.
Discontinuous series, 14
Dispersion, variations in, 79; measures of, 101 ff.; comparison of measures of, 113 ff.
Distribution, probable errors of certain constants for a normal, 237 ff. *See also* Binomial distribution, Cumulative frequency distribution, Simple frequency distribution

Efficiency, teaching, 15
Enumeration in problems, 14
Errors, absolute and relative, 65 f.; biased and unbiased, 66 ff.; in educational measurement, 74 ff.; response, 75; of estimate, 161 f.; sampling, in the mean, 232 ff.; probable, of the difference between two means, 235 ff.
Estimate, standard error of, 159, 161 f.
Estimation, in experimental work, 15; of teaching efficiency, 15
Exponents, laws of, 51 f.

Free-hand method in curve-fitting, 320, 323 ff.
Frequencies, probable errors of observed and percentage, 243 ff.
Frequency distribution. *See* Cumulative frequency distribution, Simple frequency distribution
Frequency polygon, 37 f.
Frequency table, computation of correlation coefficient for, 149 ff.
Function, hyperbola, 318; logarithmic growth, 318; *n*th-order parabola, 319
Functional relationships, 42 f.

Geometrical mean, 49; and geometrical series, 91 ff.
Geometrical progression, 48, 91 ff.
Grouping, correction for broad, 263 ff.; correction for fineness of, 269 f.

Harmonic mean, 95 f.
Heteroscedasticity, 186
Histogram, 37
Hollerith Machine, 20 f.

Imagination, constructive, as requisite, 3
Indexing, numerical or verbal mode of, 13
Intelligence and attitude, Tulchin's data on, 268 f.
Intercorrelations, 288
Interpolation, 56 ff.
Interpretation of results, 7

Kelley's formula for adjusting reliability coefficients, 254
Kurtosis, variations in, 79

Law of Statistical Regularity for Large Numbers, 16 ff.
Least squares, method of, 321 ff., 329 ff.
Line, graph of straight, 43 f.
Linearity, tests for, 183 f., 267
Logarithms, 47 ff.; invention of, 49 f.; laws of, 52; Briggs system of, 53; four-place table of, 60 f.; use with rounded numbers, 71 ff.

McCall's method of scaling, 226 f.
Median, 27, 85 ff.; probable error of, 238
Mode, 90 f.
Moments, 338 ff.
Multiple-response scoring formula, 171

Nomography, 31
Normal probability curve. *See* Probability curve

Ogive curve, 132
Ordinates of normal curve, 209 f.
Otis Test Scores, frequency distribution of, 25; classifier for, 26 f.; cumulative frequency distribution for, 28; histogram of, 38; illustrating the median for, 86

Partial correlations, 283 ff.; of first-order, 287; of second-order, 287
Pearson's correction formula, for broad grouping, 263 ff.; for Spearman's rank coefficient, 279 f.

Pearson's formula for product-moment correlation, 259 f.
Pearson's Tables, probable error of V with, 238
Percentage frequency, probable error of, 243
Percentile curves, 131 ff.
Percentile method, 127 ff.
Percentile ranks, 136 ff.
Percentiles, definition of, 127; computation of, 128 ff.
Permutations, 191
Planning of calculations, 7
Point binomial, 196; mean and standard deviation of, 197 f.; comparison of, and normal curve, 212 ff.
Predictive value of a test, 168
Primary records, tabulation for, 6
Probability, elementary, 192 ff.
Probability curve, normal, 44 f., 204 ff.; equation of, 207 f.; area, ordinates, and deviates of, 209 ff.
Probable error, 211; of the difference between two means, 235 ff.; of certain constants for normal distribution, 237 ff.; applications of formulas of, 240 ff.; of observed and percentage frequencies, 243 ff.; of an observed proportion, 248 ff.; of biserial r, 273; of contingency coefficient, 278; of correlation from ranks, 280; of β_1 and β_2, 344
Problem, planning study of, 5
Product-moment correlation coefficient, 143 ff., 258 ff.
Professional schools and statistical method, 2
Progressions, arithmetical and geometrical, 47 f.
Proportion, probable error of, 248 ff.; standard error of, 248

Quadrants, 40 f.
Qualitative series, 14, 256 ff., 266 ff.
Quantitative series, 14, 256 ff.
Quartile deviation, 110 ff.
Quartiles, measure of skewness based on, 122
Questionnaires, 15

Ranks, correlation from, 278 ff.
Records, tabulation for primary, 6
Rectification, 323 ff.
Regression, lines of, 154 ff.; meaning of, 163; probable errors of coefficients of, 239; probable error of higher-order coefficient of, 239; equation for, 292 ff.; coefficient of, of third-order, 315
Reliability, coefficient of, 168 ff.; Spearman-Brown formula for predicting, 169 f.; Kelley's formula for adjusting coefficient of, 254
Report, writing of, 8
Residuals, 158, 320
Response error, formulas for, 250 ff.
Results, interpretation of, 7; presentation of, 7 f.
Rounded numbers, arithmetical computation with, 69 ff.

Sample, random, 18 f.
Sampling, 16 ff.
Scaling of test questions, 224 ff.
Scores, standard, 168 f.
Selection, effect of, upon correlation, 172
Series, types of, 12 ff.; quantitative and qualitative, 14, 256 ff.; classification of, 15; correlation ratio for qualitative and unordered, 266 ff.
Sheppard's corrections, 341
Significant figures, 68
Simple frequency distribution, 22 ff.; for Otis Test Scores, 25; calculation of mean from, 79 ff.
Skewness, variations in, 79; measurement of, 122 f.
Sorting by mechanical devices, 20 f.
Source material, secondary, 10 f.
Spearman's correction for attenuation, 253
Spearman's formula based on rank differences, 278
Spearman's theorem on correlation, 168
Standard deviation, 108 ff.; of point binomial, 198; probable error of, 238
Standard error of proportion, 248
Standard scores, 168; in terms of "true" scores and response error, 251 f.
Standardized tests, use of, 2
Stanford-Binet Tests, 168
Statistical method, need for, 1 f.; general requirements for, 3 ff.; procedure in dealing with problem, 5 ff.; accuracy in, 65
Statistician, capacity required for, 4 f.

Tables, presentation of results in, 7 f.; purpose of, 31 f.; construction of, 32 ff.
Tabulation, for primary records, 6; of records, 14; by mechanical devices, 20 f.
Tallying, 22
Terman Group Intelligence Tests, 120
Test units, lack of equivalence of, 74
Tests, uses of correlation in evaluating, 167 ff.; validity of, 168; scaling of questions in, 224 ff.
Transmutation formula for comparable scores, 121
"True" scores, standard scores in terms of, 251 f.

Validity, 168
Variability, measures of, 101; absolute and relative, 117
Variables, independent and dependent, 42; method of eliminating effect of, 184 ff.; partial correlation for three, 284; partial regression equations for four, 300 ff.
Variation, coefficient of, 116 ff.

Yule, G. U., 15, 275 f.